Journal of Machine Learning Research 23 (2022) 1-32 Submitted 12/21; Revised 5/22; Published 7/22

When is the Convergence Time of Langevin Algorithms Dimension Independent? A Composite Optimization Viewpoint

Yoav Freund YFREUND@UCSD.EDU
University of California, San Diego
La Jolla, CA 92093, USA

Yi-An Ma YIANMA@UCSD.EDU
University of California, San Diego
La Jolla, CA 92093, USA

Tong Zhang TONGZHANG@TONGZHANG-ML.ORG
Google Research and the Hong Kong University of Science and Technology
New York, USA and Hong Kong

Editor: Zhihua Zhang

Abstract

There has been a surge of works bridging MCMC sampling and optimization, with a specific focus on translating non-asymptotic convergence guarantees for optimization problems into the analysis of Langevin algorithms in MCMC sampling. A conspicuous distinction between the convergence analysis of Langevin sampling and that of optimization is that all known convergence rates for Langevin algorithms depend on the dimensionality of the problem, whereas the convergence rates for optimization are dimension-free for convex problems. Whether a dimension independent convergence rate can be achieved by the Langevin algorithm is thus a long-standing open problem. This paper provides an affirmative answer to this problem for the case of either Lipschitz or smooth convex functions with normal priors. By viewing Langevin algorithm as composite optimization, we develop a new analysis technique that leads to dimension independent convergence rates for such problems.

Keywords: (Stochastic gradient) Langevin algorithm, convergence rates, Markov chain Monte Carlo, composite optimization, stochastic optimization

1. Introduction

Two of the major themes in machine learning are point prediction and uncertainty quantification. Computationally, they manifest in two types of algorithms: optimization and Markov chain Monte Carlo (MCMC). While both strategies have developed relatively separately for decades, there is a recent trend in relating both strands of research and translating nonasymptotic convergence guarantees in gradient based optimization methods to those in MCMC (Dalalyan, 2017; Dalalyan and Karagulyan, 2017; Wibisono, 2018; Mangoubi and Smith, 2017; Mangoubi and Vishnoi, 2018; Bou-Rabee et al., 2018; Ma et al., 2021). In particular, the Langevin sampling algorithm (Rossky et al., 1978; Roberts and Stramer, 2002) has been shown to be a form of gradient descent on the space of probabilities (Jordan et al., 1998; Wibisono, 2018; Bernton, 2018; Durmus et al., 2019). Many convergence rates on Langevin algorithm have emerged thenceforward, based on different assumptions on the posterior distribution (e.g., Durmus and Moulines, 2017; Cheng and Bartlett, 2018; Dwivedi et al., 2018; Durmus and Moulines, 2019; Vempala and Wibisono, 2019; Ma et al., 2019; Cheng et al., 2018a; Chatterji et al., 2018; Zou and Gu, 2019, to list a few). Because of the high dimensional nature of machine learning problems, a common focus of the previous works is the dimension dependence of the convergence rates. A number of works have focused on designing more involved algorithms to improve the dimension dependence of the MCMC convergence rates (Cheng et al., 2018b; Dalalyan and Riou-Durand, 2018; Ma et al., 2021; Shen and Lee, 2019; Mou et al., 2021; Lee et al., 2018), as discussed in detail in Section 3.

Despite the extensive effort, a conspicuous distinction between the convergence analysis of Langevin sampling and that of gradient descent still remains: all known convergence rates for Langevin algorithms depend on the dimensionality of the problem, whereas the convergence rates for gradient descent are dimension-free for convex problems. This prompts us to ask:

Can Langevin algorithm achieve dimension independent convergence rate under the usual convex assumptions?

In order to answer this question formally, we make two assumptions on the negative log-likelihood function. One is that the negative log-likelihood is convex. Another is that the negative log-likelihood is either Lipschitz continuous or smooth. Such convexity and regularity assumptions on the negative log-likelihood function correspond to a number of problems arising from application, including regression tasks such as learning Bayesian generalized linear models (McCullagh and Nelder, 1989; Box and Tiao, 1992), as well as classification tasks such as inference with Bayesian logistic regression (Gelman et al., 2004), one-layered Bayesian neural network (Neal, 1996), or Bayesian support vector machine (Sollich, 2002). We also employ a known and tractable prior distribution that is strongly log-concave—often times taken to be a normal distribution—to serve as a parallel to the L_2 regularizer in gradient descent.

Under such assumptions, we answer the above highlighted question in the affirmative. In particular, we prove that a Langevin algorithm converges similarly as convex optimization for this class of problems. In the analysis, we observe that the number of gradient queries required for the algorithm to converge does not depend on the dimensionality of the problem for either the Lipschitz continuous log-likelihood or the smooth log-likelihood equipped with a ridge separable structure.

To obtain this result, we first follow recent works (Durmus et al. (2019) in particular) and formulate the posterior sampling problem as optimizing over the Kullback-Leibler (KL) divergence, which is composed of two terms: (regularized) entropy and cross entropy. We then decompose the Langevin algorithm into two steps, each optimizing one part of the objective function. With a strongly convex and tractable prior, we explicitly integrate the diffusion along the prior distribution, optimizing the regularized entropy; whereas gradient descent over the convex negative log-likelihood optimizes the cross entropy. Via analyzing an intermediate quantity in this composite optimization procedure, we achieve a tight convergence bound that corresponds to the gradient descent's convergence for convex optimization on the Euclidean space. This dimension independent convergence rates for Lipschitz continuous log-likelihood and smooth log-likelihood endowed with a ridge separable structure carry over to the stochastic versions of the Langevin algorithm.

2. Preliminaries

2.1 Two Problem Classes

We consider sampling from a posterior distribution over parameter $w \in \mathbb{R}^d$, given the data set $\mathbf{z}$:

$$p(w|\mathbf{z}) \propto p(\mathbf{z}|w)\pi(w) \propto \exp\left(-U(w)\right),$$

where the potential function U decomposes into two parts: $U(w) = \beta^{-1}\left(f(w) + g(w)\right)$.

While the formulation is general, in the machine learning setting, $f(w)$ usually corresponds to the negative log-likelihood, and $g(w)$ corresponds to the negative log-prior. The parameter β is the temperature, which often takes the value of $1/n$ in machine learning, where n is the number of training data. The key motivation to consider this decomposition is that we assume that g is "simple" so that an SDE involving g can be solved to high precision. We will take advantage of this assumption in our algorithm design.

Assumption on function g

A0 We assume that function g is m-strongly convex ($g(w) - \frac{m}{2}\|w\|^2$ is convex) [1] and can be explicitly integrated.

Assumption on function f We assume that function f is convex (Assumption A1) and consider two cases regarding its regularity.

- In the first case, we assume that function f is G-Lipschitz continuous (Assumption $\mathrm{A2_L}$).
- In the second case, we assume that function f is L-smooth (Assumption $\mathrm{A2_S}$). We then instantiate the result by endowing it with a ridge separable structure (Assumptions R1 and R2).

The first case stems from Bayesian classification problems, where one has a simple strongly log-concave prior and a log-concave and log-Lipschitz likelihood that encodes the complexity of the data. Examples include Bayesian neural networks for classification tasks (Neal, 1996), Bayesian logistic regression (Gelman et al., 2004), as well as other Bayesian classification problems (Sollich, 2002) with Gaussian or Bayesian elastic net priors. In optimization literature, this setting corresponds to the smooth-continuous composition and is frequently examined in the stochastic composite optimization context (Lan, 2012; Duchi and Ruan, 2018). The second case corresponds to the regression type problems, where the entire posterior is strongly log-concave and log-smooth. In this case, one can separate the negative log-posterior into two parts: $\beta^{-1}g(w) = \frac{\beta^{-1}m}{2}\|w\|^2$ and $\beta^{-1}f(w) = \left(-\log p(w|\mathbf{z}) - \frac{\beta^{-1}m}{2}\|w\|^2\right)$, which is convex and $\beta^{-1}L$-smooth. We therefore directly let $g(w) = \frac{m}{2}\|w\|^2$ in Section 6.

1. We also say that the density proportional to $\exp\left(-\beta^{-1}g(w)\right)$ is $\beta^{-1}m$-strongly log-concave in this case.

2.2 Objective Functional and Convergence Criteria

We take the KL divergence $\beta^{-1}Q(p)$ to be our objective functional and solve the following optimization problem:

$$p_* = \arg\min_p Q(p), \tag{1}$$
$$Q(p) = \int p(w)\ln\frac{p(w)}{p(w|\mathbf{z})}dw = \mathbb{E}_{w\sim p}\left[f(w)+g(w)+\beta\ln p(w)\right].$$

The minimizer that solves the optimization problem (1) is the posterior distribution:

$$p_*(w) \propto \exp\left(-\beta^{-1}(f(w)+g(w))\right). \tag{2}$$

We further define the entropy functional as

$$H(p) = \beta\,\mathbb{E}_{w\sim p}\ln p(w),$$

so that the objective functional decomposes into the regularized entropy plus cross entropy:

$$Q(p) = \big(H(p) + \mathbb{E}_{w\sim p}\left[g(w)\right]\big) + \mathbb{E}_{w\sim p}\left[f(w)\right].$$

With this definition of the objective function, we state that the difference in Q leads to the KL divergence.

Proposition 1 *Let p be the solution of (1), and p' be another distribution on w. We have*

$$\mathrm{KL}(p'\|p) = \beta^{-1}[Q(p') - Q(p)].$$

This result establishes that the convergence in the objective $\beta^{-1}Q(p')$ is equivalent to the convergence in KL-divergence. Therefore our analysis will focus on the convergence of $\beta^{-1}Q(p')$.

We also define the 2-Wasserstein distance between two distributions that will become useful in our analysis.

Definition 2 *Given two probability distributions $p(x)$ and $p'(y)$ on $\mathbb{R}^d$, and let $\Pi(p,p')$ be the class of distributions $q(x,y)$ on $\mathbb{R}^d\times\mathbb{R}^d$ so that the marginals $q(x)=p(x)$ and $q(y)=p'(y)$. The W_2 Wasserstein distance of p and p' is defined as*

$$W_2(p,p')^2 = \min_{q\in\Pi(p,p')}\mathbb{E}_{(x,y)\sim q}\|x-y\|_2^2.$$

A celebrated relationship between the KL-divergence and the 2-Wasserstein distance is known as the Talagrand transport-entropy inequality (Otto and Villani, 2000).

Proposition 3 *Assume that probability density p_* is $\widehat{m}$-strongly log-concave, and p' defines another distribution on $\mathbb{R}^d$. Then p_* satisfies the log-Sobolev inequality with constant $\widehat{m}/2$* [2]*, and yields the following Talagrand inequality:*

$$W_2^2(p_*,p') \le \widehat{m}^{-1}\mathrm{KL}(p_*\|p').$$

2. This fact follows from the Bakry-Emery criterion (Bakry and Emery, 1985).

Reference	Convergence Criterion	Iteration Complexity
(Durmus et al., 2019)	$W_2(\widetilde{p}_T, p)^2 \leq \epsilon$	$\widetilde{\Omega}\left(\frac{dM+\widehat{G}^2}{\widehat{m}^2\epsilon^2}\right)$
(Chatterji et al., 2020)	$W_2(\widetilde{p}_T, p)^2 \leq \epsilon$	$\widetilde{\Omega}\left(\frac{d(M+\widehat{G}^2)}{\widehat{m}^2\epsilon}\right)$
This work (Theorem 4)	$W_2(\widetilde{p}_T, p)^2 \leq \frac{1}{m}\mathrm{KL}(\widetilde{p}_T \| p) \leq \epsilon$	$\Omega\left(\frac{\widehat{G}^2}{\widehat{m}^2\epsilon}\right)$
(Cheng and Bartlett, 2018)	$\mathrm{KL}(\widetilde{p}_T \| p) \leq \epsilon$	$\widetilde{\Omega}\left(\frac{\widehat{L}^2}{\widehat{m}^2}\frac{d}{\epsilon}\right)$
This work (Theorem 12)	$\mathrm{KL}(\widetilde{p}_T \| p) \leq \epsilon$	$\Omega\left(\frac{\widehat{L}\mathrm{trace}(\widehat{H})}{\widehat{m}^2\epsilon} + \frac{U(0)}{\epsilon}\right)$

Table 1: Comparison with Previous Results on overdamped Langevin algorithm: d is the parameter dimension, M is smoothness parameter of $\beta^{-1}g(w)$, $\widehat{m}$ is strong convexity parameter of $\beta^{-1}g(w)$, $\widehat{G}$ is the Lipschitz parameter of $\beta^{-1}f(w)$, $\widehat{L}$ is the smoothness parameter of $\beta^{-1}f(w)$, and $\widehat{H}$ is an upper bound of the Hessian matrix of $\beta^{-1}f(w)$. The first three rows correspond to the $\widehat{G}$-Lipschitz continuous likelihood whereas the last two rows correspond to the $\widehat{L}$-smooth likelihood.

3. Related Works

We compare our results to those in previous work in Table 1. Some previous works have aimed to sample from posteriors of the similar kind and obtain convergence in the KL divergence or the squared 2-Wasserstein distance.

In the Lipschitz continuous case, where the negative log-likelihood is convex and $\widehat{G}$-Lipschitz continuous, composed with an $\widehat{m}$-strongly convex and M-smooth negative log-prior, the convergence rate to achieve $W_2^2(\widetilde{p}_T, p_*) \leq \epsilon$ is $\widetilde{\Omega}\left(\frac{dM+\widehat{G}^2}{\widehat{m}^2\epsilon^2}\right)$ (Corollary 22 of Durmus et al., 2019). Similarly, (Chatterji et al., 2020) uses Gaussian smoothing to obtain a convergence rate of $\widetilde{\Omega}\left(\frac{d(M+\widehat{G}^2)}{\widehat{m}^2\epsilon}\right)$ (in Theorem 3.4), which improves the dependence on accuracy ϵ. In (Mou et al., 2019), the Metropolis-adjusted Langevin algorithm is levaraged with a proximal sampling oracle to remove the polynomial dependence on the accuracy ϵ (in total variation distance) and achieve a $\widetilde{\Omega}\left(d\log(\frac{1}{\epsilon})\right)$ convergence rate for a related composite posterior distribution. Unfortunately, an additional dimension dependent factor is always introduced into the overall convergence rate. This work demonstrates that if the m-strongly convex regularizer is explicitly integrable, then the convergence rate for the Langevin algorithm to achieve $\mathrm{KL}(\widetilde{p}_T \| p_*) \leq \epsilon$ is dimension independent: $T = \Omega\left(\frac{\widehat{G}^2}{\widehat{m}\epsilon}\right)$. This is proven in Theorem 4 for the full gradient Langevin algorithm, and in Theorem 8 for the stochastic gradient Langevin algorithm. Using Proposition 3, the result implies a bound of $T = \Omega\left(\frac{\widehat{G}^2}{\widehat{m}^2\epsilon}\right)$ to achieve $W_2^2(\widetilde{p}_T, p_*) \leq \epsilon$.

In the smooth case, where the negative log-posterior U is $\widehat{m}$-strongly convex and $\widehat{L}$-smooth, the overdamped Langevin algorithm has been shown to converge in $\widetilde{\Omega}\left(\frac{\widehat{L}^2}{\widehat{m}^2}\frac{d}{\epsilon}\right)$ number of gradient queries (Dalalyan, 2017; Dalalyan and Karagulyan, 2017; Cheng and Bartlett, 2018; Durmus and Moulines, 2019; Durmus et al., 2019), while the underdamped Langevin algorithm converges in $\widetilde{\Omega}\left(\frac{\widehat{L}^{3/2}}{\widehat{m}^2}\sqrt{\frac{d}{\epsilon}}\right)$ gradient queries (Cheng et al., 2018b; Ma et al., 2021; Dalalyan and Riou-Durand, 2018), to ensure that $\mathrm{KL}(\widetilde{p}_T \| p_*) \leq \epsilon$ and $W_2^2(\widetilde{p}_T, p_*) \leq \epsilon$. Using a randomized midpoint integration method for the underdamped Langevin dynamics, this convergence rate can be reduced to $\widetilde{\Omega}\left(\frac{\widehat{L}}{\widehat{m}^{4/3}}\left(\frac{d}{\epsilon}\right)^{1/3}\right)$ for convergence in squared 2-Wasserstein distance (Shen and Lee, 2019). This paper establishes that for overdamped Langevin algorithm, the convergence rate can be sharpened to $\Omega\left(\frac{\widehat{L}\cdot\mathrm{trace}(\widehat{H})}{\widehat{m}^2\epsilon}\right)$ to achieve $\mathrm{KL}(\widetilde{p}_T \| p_*) \leq \epsilon$, where matrix $\widehat{H}$ is an upper bound for the Hessian of function U.

Previous works have also focused on the ridge separable potential functions studied in this work. There is a literature that requires incoherence conditions on the data vectors and/or high-order smoothness conditions on the component functions to achieve a $\widetilde{\Omega}\left(\left(\frac{d}{\epsilon}\right)^{1/4}\right)$ convergence rate for $W_2^2(\widetilde{p}_T, p_*) \leq \epsilon$ using Hamiltonian Monte Carlo methods (Mangoubi and Smith, 2017; Mangoubi and Vishnoi, 2018). Making further assumptions that the differential equation of the Hamiltonian dynamics is close to the span of a small number of basis functions, this bound can be improved to polynomial in $\log(d)$ (Lee et al., 2018). Another thread of work alleviates these assumptions and achieves the $\widetilde{\Omega}\left(\left(\frac{d}{\epsilon}\right)^{1/4}\right)$ convergence rate for the general ridge separable potential functions via higher order Langevin dynamics and integration schemes (Mou et al., 2021). We follow this general ridge separable setting and assume that each individual log-likelihood is smooth. Under this assumption, we demonstrate in this paper, by instantiating the bound for the general smooth case, that the Langevin algorithm converges in $\Omega\left(\frac{1}{\epsilon}\right)$ number of gradient queries to achieve $\mathrm{KL}(\widetilde{p}_T \| p_*) \leq \epsilon$ (see Corollary 13 and Corollary 17).

4. Langevin Algorithms

We consider the following variant of the Langevin Algorithm 1.

Algorithm 1: Langevin Algorithm with Prior Diffusion

Input: Initial distribution p_0 on $\mathbb{R}^d$, stepsize η_t, $\beta = 1$

Draw w_0 from p_0

for $t = 1, 2, \ldots, T$ **do**

Sample $\widetilde{w}_t$ from $\widetilde{w}_t(\eta_t)$ with the following SDE on $\mathbb{R}^d$ and initial value $\widetilde{w}_t(0) = w_{t-1}$

$$\widetilde{w}_t(\eta_t) = w_{t-1} - \int_0^{\eta_t} \nabla g(\widetilde{w}_t(s))ds + \sqrt{2\beta}\int_0^{\eta_t} dB_s, \tag{3}$$

where dB_s is the standard Brownian motion on $\mathbb{R}^d$.

Let

$$w_t = \widetilde{w}_t - \widetilde{\eta}_t \nabla f(\widetilde{w}_t) \tag{4}$$

end

return $\widetilde{w}_T$

In this method, we assume that the prior diffusion equation (3) can be solved efficiently. When the prior distribution is a standard normal distribution where $g(w) = \frac{m}{2}\|w\|_2^2$ on $\mathbb{R}^d$, we can instantiate equation (3) to be:

$$\text{Sample} \quad \widetilde{w}_t(\eta_t) \sim \mathcal{N}\left(e^{-m\eta_t} w_{t-1}, \frac{1 - e^{-2m\eta_t}}{m}\beta \mathrm{I}\right). \tag{5}$$

In general, the diffusion equation (3) can also be solved numerically for separable $g(w)$ of the form

$$g(w) = \sum_{j=1}^{d} g_j(w_j),$$

where $w = [w_1, \ldots, w_d]$. In this case, we only need to solve d one-dimensional problems, which are relatively simple. For example, this includes the $L_1 - L_2$ regularization arising from the Bayesian elastic net (Li and Lin, 2010),

$$g(w) = \frac{m}{2}\|w\|_2^2 + \alpha\|w\|_1,$$

among other priors that decompose coordinate-wise.

We will also consider the stochastic version of Algorithm 1, the stochastic gradient Langevin dynamics (SGLD) method, with a strongly convex function $g(w)$. Assume that function f decomposes into $f(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$. Let D be the distribution over the dataset Ω such that expectation over it provides the unbiased estimate of the full gradient: $\mathbb{E}_{z\sim D}\nabla_w \ell(w, z) = \nabla f(w)$. Then the new algorithm takes the following form and can be instantiated in the same way as Algorithm 1.

Algorithm 2: Stochastic Gradient Langevin Algorithm with Prior Diffusion

Input: Initial distribution p_0 on $\mathbb{R}^d$, stepsize η_t, $\beta = 1/n$

Draw w_0 from p_0

for $t = 1, 2, \ldots, T$ **do**

Sample $\widetilde{w}_t$ from $\widetilde{w}_t(\eta_t)$ with the following SDE on $\mathbb{R}^d$ and initial value $\widetilde{w}_t(0) = w_{t-1}$

$$\widetilde{w}_t(\eta_t) = w_{t-1} - \int_0^{\eta_t} \nabla g(\widetilde{w}_t(s))ds + \sqrt{2\beta}\int_0^{\eta_t} dB_s, \tag{6}$$

where dB_s is the standard Brownian motion on $\mathbb{R}^d$.

Draw minibatch $\mathcal{S}$ where each $z_i \in \mathcal{S}$ are i.i.d. draws: $z_i \sim D$. Let

$$w_t = \widetilde{w}_t - \widetilde{\eta}_t \frac{1}{|\mathcal{S}|}\sum_{z_i \in \mathcal{S}} \nabla_w \ell(\widetilde{w}_t, z_i). \tag{7}$$

end

return $\widetilde{w}_T$

This algorithm becomes the streaming SGLD method where in each iteration we take one data point $z \sim D$.

In the analysis of Algorithm 1, we will use p_{t-1} to denote the distribution of w_{t-1}, and $\widetilde{p}_t$ to denote the distribution of $\widetilde{w}_t$, where the randomness include all random sampling in the algorithm. When using samples along the Markov chain to estimate expectations over function $\phi(\cdot)$, we take a weighted average, so that

$$\hat{\phi}(p) = \frac{1}{\sum_{s=1}^{T} \eta_s} \sum_{t=1}^{T} \eta_t \phi(\widetilde{w}_t),$$

which is equivalent to the expectation with respect to the weighted averaged distribution:

$$\bar{p}_T = \frac{1}{\sum_{s=1}^{T} \eta_s} \sum_{t=1}^{T} \eta_t \widetilde{p}_t.$$

Similar to stochastic optimization (Polyak and Juditsky, 1992), we prove in what follows the convergence of weighted average of the distributions $\widetilde{p}_t$ along the updates of (3) and (4) towards the posterior distribution (2).

5. Langevin Algorithms in Lipschitz Convex Case

For the posterior $p(w|\mathbf{z}) \propto \left(-\beta^{-1}(f(w)+g(w))\right)$, we assume that function f satisfies the following two conditions common to convex analysis.

Assumptions for the Lipschitz Convex Case:

A1 Function $f : \mathbb{R}^d \to \mathbb{R}$ is convex.

A2$_{\mathrm{L}}$ Function f is G-Lipschitz continuous on $\mathbb{R}^d$: $\|\nabla f(w)\|_2 \leq G$.

We also assume that function $g : \mathbb{R}^d \to \mathbb{R}$ is m-strongly convex. Note that we have assumed that the gradient of function f exists but have not assumed that function f is smooth.

5.1 Full Gradient Langevin Algorithm Convergence in Lipschitz Convex Case

Our main result for Full Gradient Langevin Algorithm in the case that f is Lipshitz can be stated as follows.

Theorem 4 *Assume that function f satisfies the convex and Lipschitz continuous Assumptions* A1 *and* A2$_{\mathrm{L}}$. *Further assume that function $g(w)$ satisfies Assumption* A0. *Then for $\widetilde{p}_T$ following the Langevin Algorithm 1, it satisfies (for $\widetilde{\eta}_t = \left(1-e^{-m\eta_t}\right)/m = 2/\left(m(t+2)\right)$):*

$$\sum_{t=1}^{T} \frac{1+0.5t}{T+0.25T(T+1)} \beta^{-1}[Q(\widetilde{p}_t) - Q(p_*)] \leq \frac{5G^2}{\beta m T}.$$

By the convexity of the KL divergence, $\beta^{-1}Q$, this leads to the convergence rate of

$$T = \frac{5G^2}{\beta m \epsilon},$$

for the averaged distribution $\bar{p}_T = \sum_{t=1}^{T} \frac{1+0.5t}{T+0.25T(T+1)} \widetilde{p}_t$ to convergence to ϵ accuracy in the KL-divergence.

We devote the rest of this section to prove Theorem 4.

Proof [Proof of Theorem 4] We take a composite optimization approach and analyze the convergence of the Langevin algorithm in two steps. First we characterize the decrease of the regularized entropy $\mathbb{E}_{w\sim p}\left[g(w)+H(p)\right]$ along the diffusion step (3).

Lemma 5 (For Regularized Entropy) *We generalize Lemma 5 of (Durmus et al., 2019) and have for $\widetilde{p}_t$ being the density of $\widetilde{w}_t$ following equation* (3) *and p being another probability density,*

$$\frac{2}{m}(1-e^{-m\eta_t})\left(\mathbb{E}_{w\sim\widetilde{p}_t}[g(w)+H(\widetilde{p}_t)]-\mathbb{E}_{w\sim p}[g(w)+H(p)]\right)\leq e^{-m\eta_t}W_2^2(p_{t-1},p)-W_2^2(\widetilde{p}_t,p),$$

where m is the strong convexity of $g(w)$.

We then capture the decrease of the cross entropy $\mathbb{E}_{w\sim p}\left[f(w)\right]$ along the gradient descent step (4). This result follows and parallels the standard convergence analysis of gradient descent (see Zinkevich, 2003; Zhang, 2004, for example).

Lemma 6 *Given probability density p on $\mathbb{R}^d$. Define*

$$f(p)=\mathbb{E}_{w\sim p}f(w),$$

then we have for p_t being the density of w_t following equation (4)*:*

$$2\widetilde{\eta}_t[f(\widetilde{p}_t)-f(p)]\leq W_2^2(\widetilde{p}_t,p)-W_2^2(p_t,p)+\widetilde{\eta}_t^2G^2.$$

We then combine the two steps to prove the overall convergence rate for the Langevin algorithm. It is worth noting that by aligning the diffusion step (3) and the gradient descent step (4) at $\widetilde{p}_t$, we combine $\mathbb{E}_{w\sim\widetilde{p}_t}[g(w)+H(\widetilde{p}_t)]$ with $f(\widetilde{p}_t)$ and cancel out $W_2^2(\widetilde{p}_t,p)$ perfectly and achieve the same convergence rate as that of stochastic gradient descent in optimization.

Proposition 7 *Set $\widetilde{\eta}_t=(1-e^{-m\eta_t})/m=\tau\cdot(\tau/\widetilde{\eta}_0+mt)^{-1}$ for some $\tau\geq 1$ and $\widetilde{\eta}_0>0$. Then*

$$\sum_{t=1}^{T}\widetilde{\eta}_t^{1-\tau}[Q(\widetilde{p}_t)-Q(p)]\leq\widetilde{\eta}_0^{-\tau}W_2^2(p_0,p)+G^2\sum_{t=1}^{T}\widetilde{\eta}_t^{2-\tau}.$$

Choosing $\tau=2$ and $p=p_*$, we have

$$\sum_{t=1}^{T}\frac{1+0.5t}{T+0.25T(T+1)}\beta^{-1}[Q(\widetilde{p}_t)-Q(p_*)]\leq\frac{4}{\beta m\widetilde{\eta}_0^2T(T+1)}W_2^2(p_0,p_*)+\frac{4G^2}{\beta m(T+1)}. \quad (8)$$

The learning rate schedule of $\eta_t=1/mt$ (with $\tau=1$) was introduced to SGD analysis for strongly convex objectives in (Shalev-Shwartz et al., 2011), which leads to a similar rate as that of Proposition 7, but with an extra $\log(T)$ term than (8). The use of $\tau>1$ has been adopted in more recent literature of SGD analysis, as an effort to avoid the $\log(T)$ term (for example, see (Lacoste-Julien et al., 2012)). The resulting bound in the SGD analysis becomes identical to that of Proposition 7, and this rate is optimal for nonsmooth strongly convex optimization (Rakhlin et al., 2012). In addition, it is possible to implement for Langevin algorithm a similar scheme using moving averaging, as discussed in (Shamir and Zhang, 2013).

It can be observed that taking a large step size $\widetilde{\eta}_0$ will grant rapid convergence. The largest one can take is to choose $\eta_0 = +\infty$ and consequently $\widetilde{\eta}_0 = 1/m$, leading to a learning rate schedule of $\widetilde{\eta}_t = 2/(m \cdot (t+2))$. In this case, we are effectively initializing from $\widetilde{p}_1 \propto \exp\left(-\beta^{-1} g(w)\right)$. Choosing the same $p_0 \propto \exp\left(-\beta^{-1} g(w)\right)$, we can bound the initial error $W_2^2(p_0, p_*)$ via the Talagrand inequality in Proposition 1 and the log-Sobolev inequality (Bakry and Emery, 1985; Ledoux, 2000) for the $\beta^{-1}m$-strongly log-concave distribution p_*:

$$W_2^2(p_0, p_*) \le \frac{\beta}{m} \mathrm{KL}(p_* \| p_0) \le \frac{\beta^2}{2m^2} \mathbb{E}_{p_*}\left[\left\|\nabla \log \frac{p_*}{p_0}\right\|^2\right] \le \frac{G^2}{2m^2},$$

since $\left\|\nabla \log \frac{p_*}{p_0}(w)\right\| = \left\|\beta^{-1} \nabla f(w)\right\| \le \beta^{-1} G$. Plugging this bound and $\widetilde{\eta}_0 = 1/m$ into equation (8), and noting that $T \ge 1$, we arrive at our result that

$$\sum_{t=1}^{T} \frac{1+0.5t}{T + 0.25T(T+1)} \beta^{-1} [Q(\widetilde{p}_t) - Q(p_*)] \le \frac{5G^2}{\beta m T}.$$

■

Proof [Proof of Proposition 7] We can add the inequalities in Lemma 5 and Lemma 6 to obtain:

$$\widetilde{\eta}_t [Q(\widetilde{p}_t) - Q(p)] \le e^{-m\eta_t} W_2(p_{t-1}, p)^2 - W_2(p_t, p)^2 + \widetilde{\eta}_t^2 G^2.$$

This is equivalent to

$$\widetilde{\eta}_t^{1-\tau} [Q(\widetilde{p}_t) - Q(p)] \le (1 - m\widetilde{\eta}_t) \widetilde{\eta}_t^{-\tau} W_2(p_{t-1}, p)^2 - \widetilde{\eta}_t^{-\tau} W_2(p_t, p)^2 + \widetilde{\eta}_t^{2-\tau} G^2. \tag{9}$$

We first show that

$$(1 - m\widetilde{\eta}_t) \widetilde{\eta}_t^{-\tau} \le \widetilde{\eta}_{t-1}^{-\tau}. \tag{10}$$

Let $s = t + \tau/(m\widetilde{\eta}_0) \ge 1$ for $t \ge 1$, $\tilde{\eta}_t = \tau/(ms)$ and $\tilde{\eta}_t = \tau/(m(s-1))$. Therefore (10) is equivalent to

$$(1 - \tau/s) s^\tau \le (s-1)^\tau.$$

This inequality follows from the fact that for $z = 1/s \in [0, 1]$ and $\tau \ge 1$: $\psi(z) = (1-z)^\tau$ is convex in z, and thus $(1 - \tau z) = \psi(0) + \psi'(0) z \le \psi(z) = (1-z)^\tau$.

By combining (9) and (10), we obtain

$$\widetilde{\eta}_t^{1-\tau} [Q(\widetilde{p}_t) - Q(p)] \le \widetilde{\eta}_{t-1}^{-\tau} W_2(p_{t-1}, p)^2 - \widetilde{\eta}_t^{-\tau} W_2(p_t, p)^2 + \widetilde{\eta}_t^{2-\tau} G^2.$$

By summing over $t = 1$ to $t = T$, we obtain the bound. ■

5.2 Streaming SGLD Convergence in Lipschitz Convex Case

To analyze the streaming stochastic gradient Langevin algorithm, we assume that function f decomposes:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i) = \mathbb{E}_{z\sim D}[\ell(w, z)],$$

where D is the distribution over the data samples. In this case, we modify Assumption A2_{L} and assume that the individual log-likelihood satisfies the Lipschitz condition.

Assumptions on individual loss ℓ

A$2_{\mathrm{L}}^{\mathrm{SG}}$ Function ℓ is G_ℓ-Lipschitz continuous on $\mathbb{R}^d$: $\|\nabla \ell(w, z)\|_2 \le G_\ell$, $\forall z \in \Omega$.

In the case that $\ell(w, z)$ is Lipschitz, our main result for SGLD is the following counterpart of Theorem 4.

Theorem 8 *Assume that function f satisfies the convex assumption* A1 *and the Lipschitz continuous assumption for the individual log-likelihood* A$2_{\mathrm{L}}^{\mathrm{SG}}$. *Further assume that function $g(w)$ satisfies Assumption* A0. *Then for $\widetilde{p}_T$ following the streaming SGLD Algorithm 2, it satisfies (for $\widetilde{\eta}_t = \left(1 - e^{-m\eta_t}\right)/m = 2/\left(m(t+2)\right)$):*

$$\sum_{t=1}^{T} \frac{1+0.5t}{T+0.25T(T+1)} \beta^{-1}[Q(\widetilde{p}_t) - Q(p_*)] \le \frac{5G_\ell^2}{\beta m T}.$$

leading to the convergence rate of

$$T = \frac{5G_\ell^2}{\beta m \epsilon},$$

for the averaged distribution $\bar{p}_T = \sum_{t=1}^{T} \frac{1+0.5t}{T+0.25T(T+1)}\widetilde{p}_t$ to convergence to ϵ accuracy in the KL-divergence.

This result corresponds to the convergence behavior of stochastic strongly convex optimization with a bounded gradient oracle (Hazan and Kale, 2014; Agarwal et al., 2012). We devote the rest of this section to prove Theorem 8.

Proof [Proof of Theorem 8] Same as in the previous section, convergence of the regularized entropy $\mathbb{E}_{w\sim p}\left[g(w)\right] + H(p)$ along equation (6) follows Lemma 5.

For the convergence of the cross entropy $\mathbb{E}_{w\sim p}\left[f(w)\right]$ along equation (7), the following Lemma follows the standard analysis of SGD.

Lemma 9 *Adopt Assumption* A$2_{\mathrm{L}}^{\mathrm{SG}}$ *that $\ell(w, z)$ is G_ℓ-Lipschitz for all $z \in \Omega$. Also adopt Assumption* A1 *that $f(w) = \mathbb{E}_{z\sim D}\ell(w, z)$ is convex. We have for all $w \in \mathbb{R}^d$:*

$$2\widetilde{\eta}_t \mathbb{E}_{z\sim D}[\ell(\widetilde{w}_t, z) - \ell(w, z)] \le \|\widetilde{w}_t - w\|_2^2 - \mathbb{E}_{w_t|\widetilde{w}_t}\|w_t - w\|_2^2 + \widetilde{\eta}_t^2 G_\ell^2. \tag{11}$$

It implies the following bound, which modifies Lemma 6.

Lemma 10 *Given any probability density q on $\mathbb{R}^d$. Define*

$$\ell(q) = \mathbb{E}_{w\sim q}\mathbb{E}_{z\sim D}\ell(w, z),$$

then we have

$$2\widetilde{\eta}_t[\ell(\widetilde{p}_t) - \ell(p)] \leq W_2(\widetilde{p}_t, p)^2 - W_2(p_t, p)^2 + \widetilde{\eta}_t^2 G_\ell^2.$$

Initializing from the prior distribution, we can follow the same proof as in Proposition 7 and obtain a similar convergence rate as in the non-stochastic case.

Proposition 11 *Set $\widetilde{\eta}_t = (1 - e^{-m\eta_t})/m = \tau \cdot (\tau/\widetilde{\eta}_0 + mt)^{-1}$ for some $\tau \geq 1$ and $\widetilde{\eta}_0 > 0$. Then*

$$\sum_{t=1}^{T} \widetilde{\eta}_t^{1-\tau}[Q(\widetilde{p}_t) - Q(p)] \leq \eta_0^{-\tau} W_2(p_0, p)^2 + G_\ell^2 \sum_{t=1}^{T} \widetilde{\eta}_t^{2-\tau}.$$

We can choose $\tau = 2$, and then for $p = p_*$, we have

$$\sum_{t=1}^{T} \frac{1 + 0.5t}{T + 0.25T(T+1)}\beta^{-1}[Q(\widetilde{p}_t) - Q(p_*)] \leq \frac{4}{\beta m \widetilde{\eta}_0^2 T(T+1)} W_2(p_0, p_*)^2 + \frac{4G_\ell^2}{\beta m(T+1)}. \tag{12}$$

Following the same steps as in the full gradient case, we arrive at the result. ■

6. Langevin Algorithms in Smooth Convex Case

For the posterior $p(w|\mathbf{z}) \propto \left(-\beta^{-1}(f(w) + g(w))\right)$, we make the following assumptions on function f.

Assumptions for the smooth convex case:

A1 Function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and positive.

A2$_\mathrm{S}$ Function f is L-Smooth on $\mathbb{R}^d$: $\|\nabla f(w) - \nabla f(w')\|_2 \leq L\|w - w'\|_2$.

We also assume that function $g : \mathbb{R}^d \to \mathbb{R}$ is m-strongly convex. Note that this is equivalent to the cases where we simply assume the entire negative log-posterior to be $\beta^{-1}m$-strongly convex and $\left(\beta^{-1}(L+m)\right)$-smooth: one can separate the negative log-posterior into two parts: $\frac{\beta^{-1}m}{2}\|w\|^2$ and $\left(-\log p(w|\mathbf{z}) - \frac{\beta^{-1}m}{2}\|w\|^2\right)$, which is convex and $\beta^{-1}L$-smooth. We therefore directly let $g(w) = \frac{m}{2}\|w\|^2$ in what follows.

6.1 Full Gradient Langevin Algorithm Convergence in Smooth Convex Case

Our main result for Full Gradient Langevin Algorithm in the case that f is smooth can be stated as follows. Compared to Theorem 4, the result of Theorem 12 is useful for loss functions such as least squares loss that are smooth but *not Lipschitz continuous.*

Theorem 12 *Assume that function f satisfies the convex and smooth Assumptions* A1 *and* A2$_{\mathrm{S}}$. *Also assume that $\nabla^2 f(w) \preceq H$. Further let function $g(w) = \frac{m}{2}\|w\|^2$, and set $\widetilde{\eta}_t = \left(1 - e^{-m\eta_t}\right)/m = 2\cdot((8L+mt)^{-1}$ with $\widetilde{\eta}_0 = 1/(4L)$. Then for $\widetilde{p}_T$ following Algorithm 1 and initializing from $p_0 \propto \exp(-\beta^{-1}g)$, it satisfies:*

$$\sum_{t=1}^{T} \frac{(4L/m)+t/2}{(4L/m)T + T(T+1)/4}\beta^{-1}[Q(\widetilde{p}_t) - Q(p_*)]$$
$$\leq \frac{64L^2}{m^2T(T+1)} \cdot \left(\frac{L}{m^2}\mathrm{trace}\,(H) + 2U(0)\right) + \frac{16}{T+1} \cdot \left(\frac{L}{m^2}\mathrm{trace}\,(H) + 2U(0)\right).$$

leading to the convergence rate of

$$T = 64 \cdot \max\left\{\frac{L \cdot \mathrm{trace}\,(H)}{m^2\epsilon}, \frac{2U(0)}{\epsilon}\right\},$$

for the averaged distribution $\bar{p}_T = \sum_{t=1}^{T} \frac{(4L/m)+0.5t}{(4L/m)T+0.25T(T+1)}\widetilde{p}_t$ to convergence to $\epsilon \leq 1$ accuracy in the KL-divergence.

Note that in the worst case, $\mathrm{trace}\,(H)$ can have dimension dependence. We discuss in the following the ridge separable case where $\mathrm{trace}\,(H)$ does not depend on the dimension d of the problem.

Ridge Separable Case Assume that function f decomposes into the following ridge-separable form:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} s_i(w^\top z_i), \tag{13}$$

We make some assumptions on the activation function s_i and the data points z_i.

Assumptions in ridge separable case

R1 $\forall i \in \{1, \ldots, n\}$, the one dimensional activation function $s_i(\cdot)$ has a bounded second derivative: $|s_i''(x)| \leq L_s$, for any $x \in \mathbb{R}$.

R2 $\forall i \in \{1, \ldots, n\}$, data point $z_i \in \mathbb{R}^d$ has a bounded norm: $\|z_i\|^2 \leq R_z$.

Assumptions R1 and R2 combines to give a smoothness constant of L_sR_z for the individual log-likelihood.

Corollary 13 *We make the convexity Assumption* A1 *on function f and let it take the ridge-separable form* (13) *(also let function $g(w) = \frac{m}{2}\|w\|^2$). Further adopt Assumptions* R1 *and* R2 *on the activation functions and the data points, respectively. Then the convergence* *rate of Algorithm 1 initializing from $p_0 \propto \exp(-\beta^{-1} g)$ (with step size $\widetilde{\eta}_t = (1 - e^{-m\eta_t})/m = 2\left(8L_s R_z + mt\right)^{-1}$) is*

$$T = 64 \cdot \max\left\{\frac{L_s^2 R_z^2}{m^2 \epsilon}, \frac{2U(0)}{\epsilon}\right\},$$

for the averaged distribution to convergence to ϵ accuracy in the KL-divergence $\beta^{-1}Q$.

Proof We first compute using the form of f that $\nabla^2 f(w) = \frac{1}{n}\sum_{i=1}^n s_i''(w^\top z_i) z_i z_i^\top$. From Assumptions R1 and R2, we know that $\nabla^2 f(w) \preceq \frac{1}{n} L_s Z Z^\top = H$, where we denote matrix $Z = (z_1, \ldots, z_n)$.

Hence the Lipschitz constant $L \le \|H\|_2 \le L_s R_z$, and

$$\operatorname{trace}(H) = L_s \cdot \frac{1}{n}\operatorname{trace}\left(ZZ^\top\right) \le L_s R_z.$$

These two facts lead to the conclusion that $L \cdot \operatorname{trace}(H) \le L_s^2 R_z^2$. Plugging the bound into Theorem 12 yields the convergence rate of

$$T = 64 \cdot \max\left\{\frac{L_s^2 R_z^2}{m^2 \epsilon}, \frac{2U(0)}{\epsilon}\right\}.$$

■

We devote the rest of this section to the proof of Theorem 12.

Proof [Proof of Theorem 12] Same as in Section 5.1, convergence of the regularized entropy $\mathbb{E}_{w\sim p}\left[g(w)\right] + H(p)$ along equation (6) follows Lemma 5.

For the decrease of the cross entropy $\mathbb{E}_{w\sim p}\left[f(w)\right]$ along the gradient descent step (4), we use the following derivation for L-smooth f. For p_t being the density of w_t following equation (4) and for p being another probability density,

$$\begin{aligned}
&2\widetilde{\eta}_t[\mathbb{E}_{w\sim\widetilde{p}_t} f(w) - \mathbb{E}_{w'\sim p} f(w')] \qquad (14)\\
\le&[W_2(\widetilde{p}_t, p)^2 - W_2(p_t, p)^2] + \eta_t^2 \mathbb{E}_{w\sim\widetilde{p}_t}\|\nabla f(w)\|_2^2\\
\le&[W_2(\widetilde{p}_t, p)^2 - W_2(p_t, p)^2] + 2\widetilde{\eta}_t^2 \mathbb{E}_{(w,w')\sim\gamma_t}\|\nabla f(w) - \nabla f(w')\|_2^2 + 2\widetilde{\eta}_t^2 \mathbb{E}_{w'\sim p}\|\nabla f(w')\|_2^2,
\end{aligned}$$

where $\gamma_t \in \Gamma_{\mathrm{opt}}(\widetilde{p}_t, p)$ is the optimal coupling between distributions with densities $\widetilde{p}_t$ and p.

With $\widetilde{\eta}_t = (1 - e^{-m\eta_t})/m$, we have

$$\begin{aligned}
2\widetilde{\eta}_t[Q(\widetilde{p}_t) - Q(p)] \le{}& (1 - m\widetilde{\eta}_t) W_2(\widetilde{p}_{t-1}, p)^2 - W_2(p_t, p)^2\\
&+ 2\widetilde{\eta}_t^2 \mathbb{E}_{(w,w')\sim\gamma_t}\|\nabla f(w) - \nabla f(w')\|_2^2 + 2\widetilde{\eta}_t^2 \mathbb{E}_{w'\sim p}\|\nabla f(w')\|_2^2. \qquad (15)
\end{aligned}$$

We also have the following lemma.

Lemma 14 *Let $\gamma_t \in \Gamma_{opt}(\widetilde{p}_t, p)$ be the optimal coupling of $\widetilde{p}_t$ and p, and let p the solution of (1). Then we have*

$$\mathbb{E}_{(w,w')\sim\gamma_t}\|\nabla f(w) - \nabla f(w')\|_2^2 \le 2L[Q(\widetilde{p}_t) - Q(p)].$$

We note that Lemma 14 and equation (15) imply that

$$\begin{aligned}&2\widetilde{\eta}_t(1-2L\widetilde{\eta}_t)[Q(\widetilde{p}_t)-Q(p_*)]\\ \leq&(1-m\widetilde{\eta}_t)W_2(p_{t-1},p_*)^2-W_2(p_t,p_*)^2+2\widetilde{\eta}_t^2\mathbb{E}_{w\sim p_*}\|\nabla f(w)\|_2^2,\end{aligned} \tag{16}$$

where p_* satisfies (2).

Next we bound the last term of equation (16) at p_*: $\mathbb{E}_{w\sim p_*}\|\nabla f(w)\|_2^2$.

Lemma 15 *Assume that*

$$\nabla^2 f(w)\preceq H,\quad g(w)=\frac{m}{2}\|w\|_2^2.$$

Let

$$w_*=\arg\min_w\left[f(w)+g(w)\right],$$

and $p_*\propto\exp(-\beta^{-1}(f(w)+g(w)))$ satisfy equation (2). Then

$$\mathbb{E}_{w\sim p_*}\left\|\nabla f(w)\right\|^2\leq\frac{2\beta L}{m}\mathrm{trace}(H)+2m^2\left\|w_*\right\|^2.$$

With these lemmas, we are ready to prove the convergence rate of the Langevin algorithm 1.

We note that similar to (10), the shrinking step size scheduling of $\widetilde{\eta}_t=\tau\cdot\left(\frac{\tau}{\widetilde{\eta}_0}+mt\right)^{-1}$ satisfies:

$$(1-m\widetilde{\eta}_t)\leq\frac{\widetilde{\eta}_t^{\tau}}{\widetilde{\eta}_{t-1}^{\tau}}.$$

Using this inequality and combining Lemma 15 and equation (16) at $p=p_*$, we obtain that

$$\begin{aligned}&2\widetilde{\eta}_t^{1-\tau}\left(1-2L\widetilde{\eta}_t\right)\left[Q(\widetilde{p}_t)-Q(p_*)\right]\\ \leq&\widetilde{\eta}_{t-1}^{-\tau}W_2(p_{t-1},p_*)^2-\widetilde{\eta}_t^{-\tau}W_2(p_t,p_*)^2+4\widetilde{\eta}_t^{2-\tau}\left[(\beta L/m)\,\mathrm{trace}\left(H\right)+m^2\left\|w_*\right\|^2\right].\end{aligned}$$

Summing over $t=1,\ldots,T$,

$$\begin{aligned}&2\sum_{t=1}^{T}\widetilde{\eta}_t^{1-\tau}\left(1-2L\widetilde{\eta}_t\right)\left[Q(\widetilde{p}_t)-Q(p_*)\right]\\ \leq&\widetilde{\eta}_0^{-\tau}W_2(p_0,p_*)^2+4\left[(\beta L/m)\,\mathrm{trace}\left(H\right)+m^2\left\|w_*\right\|^2\right]\sum_{t=1}^{T}\widetilde{\eta}_t^{2-\tau}.\end{aligned}$$

Denote $\Delta=4\left[(\beta L/m)\,\mathrm{trace}\left(H\right)+m^2\left\|w_*\right\|^2\right]$ and take $\widetilde{\eta}_t=\tau\cdot\left(\frac{\tau}{\widetilde{\eta}_0}+mt\right)^{-1}$. Since $1-2\widetilde{\eta}_tL\geq1-2\widetilde{\eta}_0L\geq0.5$, we have for $\tau=2$,

$$m\sum_{t=1}^{T}\left(1/\left(m\widetilde{\eta}_0\right)+t/2\right)\left[Q(\widetilde{p}_t)-Q(p_*)\right]\leq\frac{1}{\widetilde{\eta}_0^2}W_2(p_0,p_*)^2+\Delta\cdot T,$$

or

$$\sum_{t=1}^{T} \frac{1/\left(m\widetilde{\eta}_0\right) + t/2}{T/\left(m\widetilde{\eta}_0\right) + T(T+1)/4}[Q(\widetilde{p}_t) - Q(p_*)] \leq \frac{4}{m\widetilde{\eta}_0^2 T(T+1)} W_2(p_0, p_*)^2 + \frac{4\Delta}{m(T+1)}. \quad (17)$$

Inspired by the Lipschitz continuous case, we take $p_0(w) \propto \exp\left(-\beta^{-1} g(w)\right)$. Then by the Talagrand and log-Sobolev inequalities,

$$W_2(p_0, p_*)^2 \leq \frac{\beta}{m} \mathrm{KL}\left(p_* \| p_0\right) \leq \frac{\beta^2}{2m^2} \mathbb{E}_{p_*}\left[\left\|\nabla \log \frac{p_*}{p_0}\right\|^2\right] = \frac{\beta^2}{2m^2} \mathbb{E}_{p_*}\left[\left\|\beta^{-1} \nabla f(w)\right\|^2\right].$$

Applying Lemma 15 to the above inequality, we obtain that

$$W_2(p_0, p_*)^2 \leq \frac{1}{m^2}\left((\beta L/m)\,\mathrm{trace}\,(H) + m^2 \left\|w_*\right\|^2\right).$$

Then taking $\widetilde{\eta}_0 = \frac{1}{4}\frac{1}{L}$, we obtain that the weighted-averaged KL divergence is upper bounded:

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{1/\left(m\widetilde{\eta}_0\right) + t/2}{T/\left(m\widetilde{\eta}_0\right) + T(T+1)/4} \beta^{-1}[Q(\widetilde{p}_t) - Q(p_*)] \\
&\leq \frac{64L^2}{m^2 T(T+1)} \cdot \left(\frac{L}{m^2}\mathrm{trace}\,(H) + \beta^{-1} m \left\|w_*\right\|^2\right) + \frac{16}{T+1} \cdot \left(\frac{L}{m^2}\mathrm{trace}\,(H) + \beta^{-1} m \left\|w_*\right\|^2\right).
\end{aligned}$$

Since $L \leq \mathrm{trace}\,(H)$, $\forall \epsilon \leq 1$, the weighted-averaged KL divergence

$$\sum_{t=1}^{T} \frac{1/\left(m\widetilde{\eta}_0\right) + t/2}{T/\left(m\widetilde{\eta}_0\right) + T(T+1)/4} \beta^{-1}[Q(\widetilde{p}_t) - Q(p_*)] \leq \epsilon,$$

when we set

$$T \geq 64 \cdot \max\left\{\frac{L \cdot \mathrm{trace}\,(H)}{m^2 \epsilon}, \frac{\beta^{-1} m \left\|w_*\right\|^2}{\epsilon}\right\}.$$

Plugging in the bound that $m\left\|w_*\right\|^2 \leq 2f(0) = 2\beta U(0)$ gives the final result. ∎

6.2 SGLD Convergence in Smooth Convex Case

Similar to the Lipschitz continuous case, we assume that function f decomposes:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i) = \mathbb{E}_{z \sim D}[\ell(w, z)],$$

where D is the distribution over the data samples. Making the following assumption, which modifies Assumption A2$_\mathrm{S}$, that the individual log-likelihood satisfies the smooth condition yields the convergence rate for the SGLD method.

Assumptions on individual loss ℓ

$\mathrm{A2}_{\mathrm{S}}^{\mathrm{SG}}$ Function ℓ is L_ℓ-smooth on $\mathbb{R}^d$: $\forall z \in \Omega$,

$$\ell(y,z) \geq \ell(x,z) + \nabla\ell(x,z)^\top (y-x) + \frac{1}{2L_\ell}\left\|\nabla\ell(y,z) - \nabla\ell(x,z)\right\|^2.$$

$\mathrm{A3}_{\mathrm{S}}^{\mathrm{SG}}$ The stochastic gradient variance at the mode w_* is bounded:

$$\mathbb{E}_{z\sim D}\left[\left\|\nabla\ell(w_*,z) - \nabla f(w_*)\right\|^2\right] \leq b^2.$$

Assumption $\mathrm{A2}_{\mathrm{S}}^{\mathrm{SG}}$ ensures that the stochastic estimates of f are L_ℓ smooth.

Under the above assumptions, we obtain in what follows the convergence rate for the SGLD method with minibatch size $|\mathcal{S}|$. This result is the counterpart of its full gradient version in Theorem 12.

Theorem 16 *We make the convexity Assumptions* A1 *on function f and the regularity Assumptions* $\mathrm{A2}_{\mathrm{S}}^{\mathrm{SG}}$ *and* $\mathrm{A3}_{\mathrm{S}}^{\mathrm{SG}}$ *on its components ℓ. Also assume that $\nabla^2 f(w) \preceq H$. Let function $g(w) = \frac{m}{2}\|w\|^2$. Then taking $\widetilde{\eta}_t = \left(1-e^{-m\eta_t}\right)/m = 2\cdot\left(8L_sR_z + mt\right)^{-1}$, the convergence rate of the SGLD Algorithm 2 initializing from $p_0 \propto \exp(-\beta^{-1}g)$ is*

$$T = \Omega\left(\max\left\{\frac{L_\ell \mathrm{trace}(H)}{m^2\epsilon}, \frac{U(0)}{\epsilon}, \frac{1}{|\mathcal{S}|}\frac{b^2}{\beta m\epsilon}\right\}\right),$$

to achieve an accuracy of

$$\sum_{t=1}^{T}\frac{1/\left(m\widetilde{\eta}_0\right) + t/2}{T/\left(m\widetilde{\eta}_0\right) + T(T+1)/4}\beta^{-1}[Q(\widetilde{p}_t) - Q(p_*)] \leq \epsilon.$$

Comparing with the full gradient case, the last term corresponds to the strongly convex stochastic optimization with unbounded gradient oracle (Ghadimi and Lan, 2012).

Ridge Separable Case Assume that the individual component ℓ take the following form so that function f becomes ridge-separable:

$$\ell(w,z_i) = s_i(w^\top z_i). \tag{18}$$

To ensure bounded stochastic gradient variance at the mode of the posterior, we additionally assume that at the mode w_*, the derivatives of the activation functions are bounded.

Assumption in ridge separable case on bounded variance

$\mathrm{R3}^{\mathrm{SG}}$ $\exists b_s > 0$, so that $\left|s_i'(w_*^\top z_i)\right| \leq b_s$, $\forall i \in \{1,\ldots,n\}$, where $w_* = \arg\min_w\left[f(w) + g(w)\right]$.

Assumption $\mathrm{R3}^{\mathrm{SG}}$ ensures that the stochastic gradient variance is bounded at the mode. Then we have the following corollary instantiating Theorem 16.

Corollary 17 *We make the convexity Assumption* A1 *on function f and let it take the ridge-separable form* (13) *(also let function $g(w) = \frac{m}{2}\|w\|^2$). Further adopt Assumptions* R1, R2*, and* R3$^{\mathrm{SG}}$*. Then taking $\widetilde{\eta}_t = \left(1 - e^{-m\eta_t}\right)/m = 2 \cdot (8L_sR_z + mt)^{-1}$, the convergence rate of Algorithm 2 initializing from $p_0 \propto \exp(-\beta^{-1}g)$ is*

$$T = \Omega\left(\max\left\{\frac{L_s^2R_z^2}{m^2\epsilon}, \frac{U(0)}{\epsilon}, \frac{n}{|\mathcal{S}|}\frac{R_zb_s^2}{m\epsilon}\right\}\right),$$

to achieve an accuracy of

$$\sum_{t=1}^{T} \frac{1/\left(m\widetilde{\eta}_0\right) + t/2}{T/\left(m\widetilde{\eta}_0\right) + T(T+1)/4}\beta^{-1}[Q(\widetilde{p}_t) - Q(p_*)] \le \epsilon.$$

Proof [Proof of Corollary 17] Since function ℓ takes form (18), we can compute that $\nabla^2\ell(w, z_i) = s_i(w^\top z_i)z_iz_i^\top$. Using Assumptions R1 and R2, the smoothness $L_\ell = L_sR_z$. Same as in Corollary 13, we know that $\nabla^2 f(w) \preceq \frac{1}{n}L_sZZ^\top = H$. Therefore,

$$\operatorname{trace}(H) = L_s \cdot \frac{1}{n}\operatorname{trace}\left(ZZ^\top\right) \le L_sR_z,$$

leading to the fact that $L_\ell \cdot \operatorname{trace}(H) \le L_s^2R_z^2$.

For the stochastic gradient bound b at w_*, we apply Assumptions R2 and R3$^{\mathrm{SG}}$ to obtain

$$\|\nabla\ell(w_*, z_i) - \nabla\ell(w_*, z_j)\| = \left\|s_i'(w_*^\top z_i)z_i - s_j'(w_*^\top z_j)z_j\right\| \le 2\sqrt{R_z}b_s.$$

We thus have

$$\|\nabla\ell(w_*, z) - \nabla f(w_*)\| = \left\|\nabla\ell(w, z_i) - \frac{1}{n}\sum_{j=1}^{n}\nabla\ell(w, z_j)\right\| \le 2\sqrt{R_z}b_s,$$

leading to the fact that

$$\mathbb{E}_{z\sim D}\left[\|\nabla\ell(w_*, z) - \nabla f(w_*)\|^2\right] \le 4R_zb_s^2.$$

Therefore, the stochastic gradient variance bound in Assumption A3$_{\mathrm{S}}^{\mathrm{SG}}$, $b = 2\sqrt{R_z}b_s$. Plugging these bounds into Theorem 16 proves the corollary. ■

We devote the rest of this section to the proof of Theorem 16.

Proof [Proof of Theorem 16] We first note that because each $\ell(\cdot, z_i)$ is L_ℓ-smooth, the stochastic estimate of function f,

$$\widetilde{f}(w, \mathcal{S}) = \frac{1}{|\mathcal{S}|}\sum_{z_i\in\mathcal{S}}\ell(w, z_i) \tag{19}$$

is L_ℓ-smooth:

$$\begin{aligned}
\left\| \frac{1}{|\mathcal{S}|} \sum_{z_i \in \mathcal{S}} \nabla \ell(y, z_i) - \nabla \ell(x, z_i) \right\|^2 &\leq \frac{1}{|\mathcal{S}|^2} \left(\sum_{z_i \in \mathcal{S}} \| \nabla \ell(y, z_i) - \nabla \ell(x, z_i) \| \right)^2 \\
&\leq \frac{1}{|\mathcal{S}|} \sum_{z_i \in \mathcal{S}} \| \nabla \ell(y, z_i) - \nabla \ell(x, z_i) \|^2 \\
&\leq \frac{2L_\ell}{|\mathcal{S}|} \sum_{z_i \in \mathcal{S}} \Big(\ell(y, z_i) - \ell(x, z_i) - \nabla \ell(x, z_i)^\top (y - x) \Big) \\
&= 2L_\ell \left(\widetilde{f}(y) - \widetilde{f}(x) - \nabla \widetilde{f}(x)^\top (y - x) \right).
\end{aligned}$$

We thereby invoke the next lemma.

Lemma 18 *Assume that function f is convex, and that its stochastic estimate $\widetilde{f}$ is L_ℓ-smooth. Then*

$$\begin{aligned}
W_2^2(p_t, p) \leq\ & W_2^2(\widetilde{p}_t, p) - 2\widetilde{\eta}_t \left(f(\widetilde{p}_t) - f(p) \right) \\
&+ \widetilde{\eta}_t^2 \left(4L_\ell [Q(\widetilde{p}_t) - Q(p)] + 2\mathbb{E}_{(\widetilde{w}_t, w') \sim \gamma_t} \left[\mathbb{E}_{\mathcal{S}} \left\| \nabla \widetilde{f}(w', \mathcal{S}) \right\|^2 \right] \right),
\end{aligned}$$

where $f(q) = \mathbb{E}_{w \sim q} f(w)$, and $\gamma_t \in \Gamma_{\mathrm{opt}}(\widetilde{p}_t, p)$ is the optimal coupling between $\widetilde{p}_t$ and p.

Taking $\widetilde{\eta}_t = \left(1 - e^{-m\eta_t}\right)/m$ and combining Lemma 5 and Lemma 18, we obtain that

$$\begin{aligned}
&2\widetilde{\eta}_t \left(1 - 2\widetilde{\eta}_t L_\ell\right) \left(Q(\widetilde{p}_t) - Q(p)\right) \\
\leq & e^{-m\eta_t} W_2^2(p_{t-1}, p) - W_2^2(p_t, p) + 2\widetilde{\eta}_t^2 \mathbb{E}_{(\widetilde{w}_t, w') \sim \gamma_t} \left[\mathbb{E}_{\mathcal{S}} \left\| \nabla \widetilde{f}(w', \mathcal{S}) \right\|^2 \right].
\end{aligned} \tag{20}$$

We then adapt Lemma 15 to the stochastic gradient method.

Lemma 19 (Stochastic Gradient Counterpart of Lemma 15) *Assume that*

$$\nabla^2 f(w) \preceq H, \quad g(w) = \frac{m}{2} \|w\|_2^2.$$

Let

$$w_* = \arg\min_w \left[f(w) + g(w) \right],$$

and p be the solution of (2). Then for L_ℓ-smooth function $\widetilde{f}$ defined in (19), at $p = p_*$ and consequently $\gamma_t \in \Gamma_{\mathrm{opt}}(\widetilde{p}_t, p_*)$,

$$\begin{aligned}
&\mathbb{E}_{(\widetilde{w}_t, w') \sim \gamma_t} \left[\mathbb{E}_{\mathcal{S}} \left\| \nabla \widetilde{f}(w', \mathcal{S}) \right\|^2 \right] \\
\leq & \frac{2\beta L_\ell}{m} \mathrm{trace}(H) + 2m^2 \left\| w_* \right\|^2 + 2\mathbb{E}_{\mathcal{S}} \left\| \nabla \widetilde{f}(w_*, \mathcal{S}) - \nabla f(w_*) \right\|^2.
\end{aligned}$$

For the last piece of information, we establish the variance of the stochastic gradient at the mode, $\mathbb{E}_{\mathcal{S}} \left\| \nabla \widetilde{f}(w_*, \mathcal{S}) - \nabla f(w_*) \right\|_2^2$. For samples z_i that are i.i.d. draws from the data set and are unbiased estimators of $\nabla f(w_*) = \frac{1}{n}\sum_{j=1}^n \nabla \ell(w_*, z_j)$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}} \left[\left\| \sum_{z_i \in \mathcal{S}} \left(\nabla \ell(w_*, z_i) - \frac{1}{n}\sum_{j=1}^n \nabla \ell(w_*, z_j) \right) \right\|^2 \right] \\
=& |\mathcal{S}| \cdot \mathbb{E}_{z\sim D} \left[\left\| \nabla \ell(w_*, z) - \frac{1}{n}\sum_{j=1}^n \nabla \ell(w_*, z_j) \right\|^2 \right] \leq |\mathcal{S}| \cdot b^2,
\end{aligned}$$

Leading to the bound that

$$\mathbb{E}_{\mathcal{S}} \left\| \nabla \widetilde{f}(w_*, \mathcal{S}) - \nabla f(w_*) \right\|_2^2 = \frac{1}{|\mathcal{S}|^2} \mathbb{E}_{\mathcal{S}} \left[\left\| \sum_{z_i \in \mathcal{S}} \left(\nabla \ell(w_*, z_i) - \frac{1}{n}\sum_{j=1}^n \nabla \ell(w_*, z_j) \right) \right\|^2 \right] \leq \frac{b^2}{|\mathcal{S}|}.$$

Plugging this result and Lemma 19 into equation (20) at $p = p_*$, we obtain the final bound that

$$\begin{aligned}
2\widetilde{\eta}_t \left(1 - 2\widetilde{\eta}_t L_s R_z\right) \left(Q(\widetilde{p}_t) - Q(p_*)\right) \leq& \left(1 - m\widetilde{\eta}_t\right) W_2^2(p_{t-1}, p_*) - W_2^2(p_t, p_*) \\
&+ 4\widetilde{\eta}_t^2 \left(\beta \frac{L_\ell}{m} \mathrm{trace}(H) + m^2 \left\|w_*\right\|^2 + \frac{b^2}{|\mathcal{S}|} \right).
\end{aligned}$$

This leads to a convergence rate, similar to the full gradient case, of

$$T = \Omega\left(\max\left(\frac{L_\ell \mathrm{trace}(H)}{m^2 \epsilon}, \frac{\beta^{-1} m \left\|w_*\right\|^2}{\epsilon}, \frac{\beta^{-1}}{|\mathcal{S}|}\frac{b^2}{m\epsilon} \right) \right),$$

so that the weighted-averaged KL divergence:

$$\sum_{t=1}^T \frac{1/\left(m\widetilde{\eta}_0\right) + t/2}{T/\left(m\widetilde{\eta}_0\right) + T(T+1)/4} \beta^{-1} [Q(\widetilde{p}_t) - Q(p_*)] \leq \epsilon.$$

Since $m \left\|w_*\right\|^2 \leq 2f(0) = 2\beta U(0)$,

$$T = \Omega\left(\max\left(\frac{L_\ell \mathrm{trace}(H)}{m^2 \epsilon}, \frac{U(0)}{\epsilon}, \frac{1}{|\mathcal{S}|}\frac{b^2}{\beta m\epsilon} \right) \right).$$

■

We compare our analysis with some of the existing works in the smooth case to shed light on the intuition of our approach. In many previous works (e.g., Dalalyan, 2017; Dalalyan and Karagulyan, 2017; Durmus and Moulines, 2019; Cheng and Bartlett, 2018; Ma et al., 2019), the Langevin algorithm updates according to

$$\begin{aligned}
w_t &= w_{t-1} - \eta_t \nabla U(w_{t-1}) + \sqrt{2} B_{\eta_t} \\
&= w_{t-1} - \int_0^{\eta_t} \nabla U(w_{t-1}) ds + \sqrt{2} \int_0^{\eta_t} dB_s, \qquad (21)
\end{aligned}$$

where we denote $U(w) = \beta^{-1}(f(w) + g(w))$. It is analyzed via comparing against the following continuous diffusion process

$$w_t(\eta_t) = w_{t-1} - \int_0^{\eta_t} \nabla U(w_t(s))ds + \sqrt{2}\int_0^{\eta_t} dB_s; \qquad w_t(0) = w_{t-1} \tag{22}$$

during the t-th update. The diffusion process (22) converges exponentially, while the Langevin algorithm (21) contains discretization error, posing restriction on the step sizes. This discretization error is caused by the gradient ∇f being evaluated at different positions (at w_{t-1} in equation (21) versus at $w_t(s)$ in equation (22)). The discrepancy leads to the following bound in the smooth case: $\mathbb{E}\left\|\nabla U(w_t(s)) - \nabla U(w_{t-1})\right\|^2 \leq L^2/\beta^2 \cdot \mathbb{E}\left\|w_t(s) - w_{t-1}\right\|^2$. One can observe that the difference $\mathbb{E}\left\|w_t(s) - w_{t-1}\right\|^2$ contains a component $\mathbb{E}\left\|B_s\right\|^2$ that is the variance of a standard normal random variable, contributing to the dimension dependence arising from the existing analyses.

From a gradient flow perspective, the Langevin algorithm (21) dictates that the distribution p_s of random variable $w_t(s)$ follows the transport of probability mass along the vector flow: $-\nabla \ln \frac{p_s(w_t(s))}{p_*(w_{t-1})}$, when $\nabla \ln \frac{p_s(w)}{p_*(w)}$ is the strong subdifferential of the KL divergence, $\mathrm{KL}(p_s \| p_*) = \beta^{-1} Q(p_s)$. It can be observed that the numerator and the denominator of the strong subdifferential of $\mathrm{KL}(p_s \| p_*)$ are evaluated at two different positions, $w_t(s)$ and w_{t-1}. This discrepancy of gradient evaluation suggests that we should split the objective functional Q into two parts and employ a composite optimization perspective. In this approach, the tight analysis hinges upon aligning the left-hand-side of both Lemma 5 and equation (14) (or Lemma 6 in the Lipschitz continuous case) at the same intermediate variable $\widetilde{w}_t$ and its associated probability $\widetilde{p}_t$. If we only focus on the output of the algorithm, w_t, the two terms $\mathbb{E}_{w\sim p}\left[g(w) + H(p)\right]$ and $\mathbb{E}_{w\sim p}\left[f(w)\right]$ will be evaluated at different distributions. This leads to an extra suboptimal dimension dependent term for every iteration.

7. Conclusion

This paper investigated the convergence of Langevin algorithms with strongly log-concave posteriors. We assume that the strongly log-concave posterior can be decomposed into two parts, with one part being simple and explicitly integrable with respect to the underlying SDE. This is analogous to the situation of proximal gradient methods in convex optimization. Using a new analysis technique which mimics the corresponding analysis of convex optimization, we obtain convergence results for Langenvin algorithms that are independent of dimension, both for Lipschitz and for a large class of smooth convex problems in machine learning. Our result addresses a long-standing puzzle with respect to the convergence of the Langevin algorithms. We note that the current work focused on the standard Langevin algorithm, and the resulting convergence rate in terms of ϵ dependency is inferior to the best known results leveraging underdamped or even higher order Langevin dynamics such as (Cheng et al., 2018b; Dalalyan and Riou-Durand, 2018; Shen and Lee, 2019; Ma et al., 2021; Mou et al., 2021), which corresponds to accelerated methods in optimization. It thus remains open to investigate whether dimension independent bounds can be combined with these accelerated methods to improve ϵ dependence as well as condition number dependence.

8. Proofs of the Supporting Lemmas

8.1 Proofs of Lemmas in the Lipschitz Continuous Case

8.1.1 Proof of Lemma 5

Before proving Lemma 5, we first state a result in (Theorem 23.9 of Villani, 2009) that establishes the strong subdifferential of the Wasserstein-2 distance.

Lemma 20 *Assume that $\mu_t, \hat{\mu}_t$ solve the following continuity equations*

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\xi_t \mu_t) = 0, \qquad \frac{\partial \hat{\mu}_t}{\partial t} + \nabla \cdot (\hat{\xi}_t \hat{\mu}_t) = 0.$$

Then

$$\frac{1}{2}\frac{d}{dt} W_2^2(\mu_t, \hat{\mu}_t) = -\int \left\langle \tilde{\nabla}\psi_t, \xi_t \right\rangle d\mu_t - \int \left\langle \tilde{\nabla}\hat{\psi}_t, \hat{\xi}_t \right\rangle d\hat{\mu}_t,$$

where ψ_t and $\hat{\psi}_t$ are the optimal transport vector fields:

$$\exp(\tilde{\nabla}\psi_t)_{\#}\mu_t = \hat{\mu}_t, \qquad \exp(\tilde{\nabla}\hat{\psi}_t)_{\#}\hat{\mu}_t = \mu_t.$$

Writing p_t and $\hat{p}_t$ as the density functions of μ_t and $\hat{\mu}_t$, we take $\xi_t = -\beta\nabla \log p_t - \nabla g$ and $\hat{\xi}_t = 0$ so that μ_t follows the Fokker-Planck equation equation associated with process (3) and $\hat{\mu}_t = \nu$ is a constant measure. This leads to the following equation

$$\frac{1}{2}\frac{d}{ds} W_2^2(\mu_s, \nu) = \int \left\langle \beta\nabla \log p_s + \nabla g, (\tilde{\nabla}\psi)_{\mu_s}^{\nu} \right\rangle d\mu_s.$$

For μ being the probability measure associated with its density p, define relative entropy $\beta^{-1}F(\mu)$, where $F(\mu) = \mathbb{E}_{w\sim p}\left[g(w)\right] + H(p)$. We can then use the fact that the relative entropy $\beta^{-1}F$ is $\beta^{-1}m$-geodesically strongly convex (see Proposition 9.3.2 of (Ambrosio et al., 2008)) to prove the following Lemma.

Lemma 21 *For p being the density of μ,*

$$F(\nu) - F(\mu) - \frac{m}{2} W_2^2(\mu, \nu) \geq \int \left\langle \beta\nabla \log p + \nabla g, (\tilde{\nabla}\psi)_{\mu}^{\nu} \right\rangle d\mu.$$

Proof Let μ_t be the geodesic between μ and ν. $\beta^{-1}m$-geodesic strong convexity of $\beta^{-1}F$ states that (see Proposition 9.3.2 of (Ambrosio et al., 2008)):

$$\beta^{-1}F(\mu_t) \leq t\beta^{-1}F(\nu) + (1-t)\beta^{-1}F(\mu) - \frac{\beta^{-1}m}{2} t(1-t) W_2^2(\mu, \nu),$$

and consequently

$$\frac{F(\mu_t) - F(\mu)}{t} \leq F(\nu) - F(\mu) - \frac{m}{2}(1-t) W_2^2(\mu, \nu).$$

By the definition of subdifferential (c.f. Villani, 2009, Theorem 23.14) we also have along the diffusion process defined by equation (3):

$$\liminf_{t\downarrow 0} \frac{F(\mu_t) - F(\mu)}{t} \geq \int \left\langle \beta\nabla \log p + \nabla g, (\tilde{\nabla}\psi)_{\mu}^{\nu} \right\rangle d\mu.$$

Taking the limit of $t \to 0$, we obtain the result. ■

Proof [Proof of Lemma 5] Combining Lemma 20 and 21, we obtain that

$$\frac{1}{2}\frac{d}{ds}W_2^2(\mu_s,\nu) = \int \left\langle \beta\nabla\log p_s + \nabla g, (\tilde{\nabla}\psi)^{\nu}_{\mu_s} \right\rangle d\mu_s \le F(\nu) - F(\mu_s) - \frac{m}{2}W_2^2(\mu_s,\nu). \quad (23)$$

Along the Fokker-Planck equation associated with process (3), $\frac{d}{ds}F(\mu_s) = -\mathbb{E}_{p_s}\left[\|\beta\nabla\log p_s + \nabla g\|_2^2\right] \le 0$, meaning that $F(\mu_s)$ is monotonically decreasing. We obtain from equation (23) for $s \in [0,t]$,

$$\frac{1}{2}\frac{d}{ds}W_2^2(\mu_s,\nu) \le \sup_{s\in[0,t]}\left[F(\nu) - F(\mu_s)\right] - \frac{m}{2}W_2^2(\mu_s,\nu) = F(\nu) - F(\mu_t) - \frac{m}{2}W_2^2(\mu_s,\nu).$$

Applying the Gronwall's inequality, we arrive at the conclusion that

$$\frac{2}{m}\left(1-e^{-m\Delta t}\right)\left(F(\mu_t) - F(\nu)\right) \le e^{-m\Delta t}W_2^2(\mu_0,\nu) - W_2^2(\mu_t,\nu).$$

Taking $d\mu_t = \tilde{p}_t dx$, $d\mu_0 = p_{t-1}dx$, $d\nu = pdx$, and $\Delta t = \eta_t$ finishes the proof. ■

8.1.2 Proof of Lemma 6

Proof [Proof of Lemma 6] We first state a point-wise result along the gradient descent step (4):

$$2\eta_t\left(f(\widetilde{w}_t) - f(w)\right) \le \|\widetilde{w}_t - w\|_2^2 - \|w_t - w\|_2^2 + \eta_t^2 G^2. \quad (24)$$

This is because

$$\begin{aligned}
\|w_t - w\|_2^2 &= \|\tilde{w}_t - \eta_t\nabla f(\tilde{w}_t) - w\|_2^2 \\
&= \|\tilde{w}_t - w\|_2^2 - 2\eta_t\left\langle \nabla f(\tilde{w}_t), \tilde{w}_t - w\right\rangle + \eta_t^2\left\|\nabla f(\tilde{w}_t)\right\|_2^2 \\
&\le \|\tilde{w}_t - w\|_2^2 - 2\eta_t\left(f(\tilde{w}_t) - f(w)\right) + \eta_t^2 G^2,
\end{aligned}$$

where the last step follows from the convexity and Lipschitz continuity of f.

We then denote the measures corresponding to random variables w_t and $\tilde{w}_t$ to be: $w_t \sim \mu_t$ and $\tilde{w}_t \sim \tilde{\mu}_t$. From the definitions, we know that they have densities p_t and $\tilde{p}_t$.

Denote an optimal coupling between $\tilde{\mu}_t$ and μ (where measure μ has density p, which is the stationary distribution) to be $\gamma \in \Gamma_{opt}(\tilde{\mu}_t,\mu)$. We then take expectations over $\gamma(\tilde{w}_t,w)$ on both sides of equation (24):

$$\begin{aligned}
2\eta_t\left(f(\tilde{p}_t) - f(p)\right) &= 2\eta_t\mathbb{E}_{(\tilde{w}_t,w)\sim\gamma}\left[f(\tilde{w}_t) - f(w)\right] \\
&\le \mathbb{E}_{(\tilde{w}_t,w)\sim\gamma}\left[\|\tilde{w}_t - w\|_2^2\right] - \mathbb{E}_{(\tilde{w}_t,w)\sim\gamma}\left[\|w_t - w\|_2^2\right] + \eta_t^2 G^2 \\
&= W_2^2(\tilde{p}_t,p) - \mathbb{E}_{(\tilde{w}_t,w)\sim\gamma}\left[\|w_t - w\|_2^2\right] + \eta_t^2 G^2.
\end{aligned}$$

From the relationship $w_t = \tilde{w}_t - \eta_t \nabla f(\tilde{w}_t)$, we know that the joint distribution of (w_t, w) is $(\mathrm{id} - \eta_t \nabla f,\, \mathrm{id})_{\#}\, \gamma$. Note that $\tilde{\gamma} = (\mathrm{id} - \eta_t \nabla f,\, \mathrm{id})_{\#}\, \gamma$ also defines a coupling, and therefore

$$\begin{aligned}\mathbb{E}_{(\tilde{w}_t, w)\sim\gamma}\left[\|w_t - w\|_2^2\right] &= \mathbb{E}_{(w_t, w)\sim\tilde{\gamma}}\left[\|w_t - w\|_2^2\right] \\ &\geq \inf_{\hat{\gamma}\in\Gamma(\mu_t,\mu)} \mathbb{E}_{(w_t, w)\sim\hat{\gamma}}\left[\|w_t - w\|_2^2\right] = W_2^2(p_t, p).\end{aligned}$$

Therefore,

$$2\eta_t \left(f(\tilde{p}_t) - f(p)\right) \leq W_2^2(\tilde{p}_t, p) - W_2^2(p_t, p) + \eta_t^2 G^2.$$

■

8.1.3 Proofs of Lemma 9 and 10 for the streaming SGLD algorithm 2

Proof [Proof of Lemma 9] By the definitions of w_t and $\tilde{w}_t$,

$$\begin{aligned}\|w_t - w\|_2^2 &= \|\tilde{w}_t - \eta_t \nabla\ell(\tilde{w}_t, z_t) - w\|_2^2 \\ &= \|\tilde{w}_t - w\|_2^2 - 2\eta_t \left\langle \nabla\ell(\tilde{w}_t, z_t), \tilde{w}_t - w\right\rangle + \eta_t^2 \left\|\nabla\ell(\tilde{w}_t, z_t)\right\|_2^2.\end{aligned}$$

We now take expectation with respect to z_t, conditioned on $\tilde{w}_t$, to obtain

$$\begin{aligned}\mathbb{E}_{z_t|\tilde{w}_t}\|w_t - w\|_2^2 &\leq \|\tilde{w}_t - w\|_2^2 - 2\eta_t \left\langle \nabla f(\tilde{w}_t), \tilde{w}_t - w\right\rangle + \eta_t^2 G_\ell^2 \\ &\leq \|\tilde{w}_t - w\|_2^2 - 2\eta_t \left(f(\tilde{w}_t) - f(w)\right) + \eta_t^2 G_\ell^2.\end{aligned}$$

The last step follows from the convexity of f. Therefore, the desired bound follows. ■

Proof [Proof of Lemma 10] We first denote the measures corresponding to random variables w_t and $\tilde{w}_t$ to be: $w_t \sim \mu_t$ and $\tilde{w}_t \sim \tilde{\mu}_t$. From the definitions, we know that they have densities p_t and $\tilde{p}_t$.

Denote an optimal coupling between $\tilde{\mu}_t$ and μ (where measure μ has density p, which is the stationary distribution) to be $\gamma \in \Gamma_{opt}(\tilde{\mu}_t, \mu)$. We then take expectations over $\gamma(\tilde{w}_t, w)$ on both sides of Eq. (11), $\forall z \in \Omega$:

$$\begin{aligned}&2\eta_t \mathbb{E}_{(\tilde{w}_t, w)\sim\gamma}\left[\mathbb{E}_z\left[\ell(\tilde{w}_t, z) - \ell(w, z)\right]\right] \\ &\leq \mathbb{E}_{(\tilde{w}_t, w)\sim\gamma}\left[\|\tilde{w}_t - w\|_2^2\right] - \mathbb{E}_{(\tilde{w}_t, w)\sim\gamma}\left[\mathbb{E}_{w_t}\left[\|w_t - w\|_2^2 | \widetilde{w}_t\right]\right] + \eta_t^2 G_\ell^2 \\ &= W_2^2(\tilde{p}_t, p) - \mathbb{E}_{(\tilde{w}_t, w)\sim\gamma}\left[\mathbb{E}_{w_t}\left[\|w_t - w\|_2^2 | \widetilde{w}_t\right]\right] + \eta_t^2 G_\ell^2.\end{aligned}$$

From the relationship $w_t = \tilde{w}_t - \eta_t \nabla\ell(\tilde{w}_t, z_t)$, we know that conditional on z_t, the joint distribution of (w_t, w) is $(\mathrm{id} - \eta_t \nabla\ell,\, \mathrm{id})_{\#}\, \gamma$. Note that $\tilde{\gamma} = (\mathrm{id} - \eta_t \nabla\ell,\, \mathrm{id})_{\#}\, \gamma$ also defines a coupling, and therefore

$$\begin{aligned}&\mathbb{E}_{(\tilde{w}_t, w)\sim\gamma}\left[\mathbb{E}_{w_t}\left[\|w_t - w\|_2^2 | \widetilde{w}_t\right]\right] \\ &= \mathbb{E}_{z_t}\left[\mathbb{E}_{(\tilde{w}_t, w)\sim\gamma}\left[\mathbb{E}_{w_t}\left[\|w_t - w\|_2^2 | \widetilde{w}_t, z_t\right]\right]\right] \\ &= \mathbb{E}_{z_t}\left[\mathbb{E}_{(w_t, w)\sim\tilde{\gamma}}\left[\|w_t - w\|_2^2 | z_t\right]\right] \\ &\geq \inf_{\hat{\gamma}\in\Gamma(\mu_t,\mu)} \mathbb{E}_{(w_t, w)\sim\hat{\gamma}}\left[\|w_t - w\|_2^2\right] = W_2^2(p_t, p).\end{aligned}$$

Plugging this result and the Lipschitz assumption on ℓ in, we obtain that

$$2\widetilde{\eta}_t[\ell(\widetilde{p}_t) - \ell(p)] \leq W_2(\widetilde{p}_t, p)^2 - W_2(p_t, p)^2 + \widetilde{\eta}_t^2 G_\ell^2.$$

■

8.2 Proofs of Lemmas in the Lipschitz Smooth Case

8.2.1 Proofs of Lemmas 14 and 15 for the full gradient Langevin algorithm 1

Proof [Proof of Lemma 14] By the geodesic convexity of the entropy function $H(p) = \beta\mathbb{E}_{w\sim p}[\ln p(w)]$,

$$H(\widetilde{p}_t) - H(p) \geq \beta \int \left\langle \nabla \ln p(w'), \left(T_{p'}^{p} - \mathrm{id}\right)(w')\right\rangle p(w')\, dw'.$$

where $T_p^{\widetilde{p}_t}$ is the optimal transport from p to $\widetilde{p}_t$. Using optimal coupling $\mu_t \in \Pi(\widetilde{p}, p)$,

$$H(\widetilde{p}_t) - H(p) \geq \beta\mathbb{E}_{w,w'\sim\mu_t}\left[\left\langle \nabla \ln p(w'), w - w'\right\rangle\right].$$

In addition, convexity of f and g implies that

$$\mathbb{E}_{w,w'\sim\mu_t}[g(w) - g(w')] \geq \mathbb{E}_{w,w'\sim\mu_t}\left[\left\langle \nabla g(w'), w - w'\right\rangle\right]$$

and

$$\mathbb{E}_{w,w'\sim\mu_t}[f(w) - f(w')] \geq \mathbb{E}_{w,w'\sim\mu_t}\left[\left\langle \nabla f(w'), w - w'\right\rangle\right].$$

Adding the above three inequalities, and note that the following holds point-wise

$$\beta\nabla \ln p(w') + \nabla g(w') + \nabla f(w') = 0,$$

we obtain that

$$H(\widetilde{p}_t) - H(p) + \mathbb{E}_{w,w'\sim\mu_t}[g(w) - g(w')] \geq -\mathbb{E}_{w,w'\sim\mu_t}\left[\left\langle \nabla f(w'), w - w'\right\rangle\right],$$

and that

$$Q(\widetilde{p}_t) - Q(p) \geq \mathbb{E}_{w,w'\sim\mu_t}[f(w) - f(w')] - \mathbb{E}_{w,w'\sim\mu_t}\left[\left\langle \nabla f(w'), w - w'\right\rangle\right]. \tag{25}$$

Since the potential function $f(w)$ is convex and L-smooth,

$$f(w) \geq f(w') + \nabla f(w')^\top (w - w') + \frac{1}{2L}\|\nabla f(w') - \nabla f(w)\|_2^2. \tag{26}$$

Combining equations (25) and (26), we obtain the desired bound. ■

Proof [Proof of Lemma 15] From smoothness and convexity, we have

$$\begin{aligned}\|\nabla f(w)\|^2 &\leq 2\|\nabla f(w) - \nabla f(w_*)\|^2 + 2\|\nabla f(w_*)\|^2 \\ &\leq 4L\left[f(w) - f(w_*) - \nabla f(w_*)^\top (w - w_*)\right] + 2\|\nabla f(w_*)\|^2 .\end{aligned}$$

Taking expectation over w on both sides, we obtain that

$$\mathbb{E}_{w\sim p_*}\|\nabla f(w)\|^2 \leq 4L \cdot \mathbb{E}_{w\sim p_*}\left[f(w) - f(w_*) - \nabla f(w_*)^\top (w - w_*)\right] + 2\|\nabla f(w_*)\|^2 . \quad (27)$$

We now upper bound $\mathbb{E}_{w\sim p_*}\left[f(w) - f(w_*) - \nabla f(w_*)^\top (w - w_*)\right]$. Let p_0 be the normal distribution $N(w_*, (\beta/m)I)$, and define

$$\begin{aligned}\Delta f(w) &= f(w) - f(w_*) - \nabla f(w_*)^\top (w - w_*), \\ \Delta g(w) &= g(w) - g(w_*) - \nabla g(w_*)^\top (w - w_*),\end{aligned}$$

then p_* can be expressed as

$$\begin{aligned}p_*(w) &\propto \exp(-\beta^{-1}(\Delta f(w) + \Delta g(w))) \\ &\propto \exp(-\beta^{-1}\Delta f(w) + \ln p_0(w)),\end{aligned}$$

which is the solution of

$$p_* = \arg\min_p \mathbb{E}_{w\sim p}\left[\Delta f(w) + \beta \ln \frac{p(w)}{p_0(w)}\right].$$

Therefore

$$\begin{aligned}\mathbb{E}_{w\sim p_*}\Delta f(w) &\leq \mathbb{E}_{w\sim p_*}\left[\Delta f(w) + \beta \ln \frac{p_*(w)}{p_0(w)}\right] \\ &\overset{\text{(i)}}{=} -\beta \ln \mathbb{E}_{w\sim p_0} \exp(-\beta^{-1}\Delta f(w)) \\ &\leq -\beta \ln \mathbb{E}_{w\sim p_0} \exp(-0.5\beta^{-1}(w - w_*)^\top H(w - w_*)) \\ &= 0.5\beta \ln |I + H/m| \leq 0.5\beta \mathrm{trace}(H/m), \end{aligned} \quad (28)$$

where equality (i) follows from the fact that both sides equal to $-\beta \ln Z$, where Z is the normalization constant of $\exp\left(-\beta^{-1}\Delta f(w)\right)$.

We then upper bound $\|\nabla f(w_*)\|^2$. Since w_* is the minimum of $f + g$, $\nabla f(w_*) = -\nabla g(w_*) = -mw_*$. Hence

$$\|\nabla f(w_*)\|^2 \leq m^2 \|w_*\|^2 . \quad (29)$$

Plugging inequalities (28) and (29) into inequality (27) proves the desired result that

$$\mathbb{E}_{w\sim p_*}\|\nabla f(w)\|^2 \leq \frac{2\beta L}{m}\mathrm{trace}(H) + 2m^2 \|w_*\|^2 .$$

■

8.2.2 Proofs of Lemma 18 and 19 for the SGLD algorithm 2

Proof [Proof of Lemma 18] By the definitions of w_t and $\tilde{w}_t$,

$$\begin{aligned}\|w_t - w\|_2^2 &= \left\|\tilde{w}_t - \eta_t \nabla \widetilde{f}(\tilde{w}_t, \mathcal{S}) - w\right\|_2^2 \\ &= \|\tilde{w}_t - w\|_2^2 - 2\eta_t \left\langle \nabla \widetilde{f}(\tilde{w}_t, \mathcal{S}), \tilde{w}_t - w\right\rangle + \eta_t^2 \left\|\nabla \widetilde{f}(\tilde{w}_t, \mathcal{S})\right\|_2^2.\end{aligned}$$

We now take expectation with respect to $\mathcal{S}$, to obtain

$$\begin{aligned}\mathbb{E}_{w_t|\tilde{w}_t} \|w_t - w\|_2^2 &= \|\tilde{w}_t - w\|_2^2 - 2\eta_t \left\langle \nabla f(\tilde{w}_t), \tilde{w}_t - w\right\rangle + \eta_t^2 \mathbb{E}_{\mathcal{S}} \left\|\nabla \widetilde{f}(\tilde{w}_t, \mathcal{S})\right\|_2^2 \\ &\leq \|\tilde{w}_t - w\|_2^2 - 2\eta_t \left(f(\tilde{w}_t) - f(w)\right) + \eta_t^2 \mathbb{E}_{\mathcal{S}} \left\|\nabla \widetilde{f}(\tilde{w}_t, \mathcal{S})\right\|_2^2. \end{aligned} \tag{30}$$

We then upper bound $\mathbb{E}_{\mathcal{S}} \left\|\nabla \widetilde{f}(\tilde{w}_t, \mathcal{S})\right\|_2^2$ by introducing variable w' that is distributed according to p and couples optimally with the law of $\tilde{w}_t$:

$$\mathbb{E}_{\mathcal{S}} \left\|\nabla \widetilde{f}(\tilde{w}_t, \mathcal{S})\right\|_2^2 \leq 2\mathbb{E}_{\mathcal{S}} \left\|\nabla \widetilde{f}(\tilde{w}_t, \mathcal{S}) - \nabla \widetilde{f}(w', \mathcal{S})\right\|^2 + 2\mathbb{E}_{\mathcal{S}} \left\|\nabla \widetilde{f}(w', \mathcal{S})\right\|^2. \tag{31}$$

For function $\widetilde{f}$ being L_ℓ-smooth,

$$\widetilde{f}(\tilde{w}_t, \mathcal{S}) \geq \widetilde{f}(w', \mathcal{S}) + \nabla \widetilde{f}(w', \mathcal{S})^\top (w - w') + \frac{1}{2L_\ell} \|\nabla \widetilde{f}(w', \mathcal{S}) - \nabla \widetilde{f}(\tilde{w}_t, \mathcal{S})\|_2^2.$$

Taking expectation over the randomness of minibatch assignment $\mathcal{S}$ on both sides leads to the fact that

$$f(\tilde{w}_t) \geq f(w') + \nabla f(w')^\top (w - w') + \frac{1}{2L_\ell} \mathbb{E}_{\mathcal{S}} \|\nabla \widetilde{f}(w', \mathcal{S}) - \nabla \widetilde{f}(\tilde{w}_t, \mathcal{S})\|_2^2.$$

Combining this equation with equation (25), we adapt Lemma 14 to the stochastic gradient method:

$$\mathbb{E}_{(\tilde{w}_t, w') \sim \mu_t} \left[\mathbb{E}_{\mathcal{S}} \|\nabla \widetilde{f}(w', \mathcal{S}) - \nabla \widetilde{f}(\tilde{w}_t, \mathcal{S})\|_2^2\right] \leq 2L_\ell [Q(\widetilde{p}_t) - Q(p)].$$

Applying this result to equation (31) and taking expectation of $(\tilde{w}_t, w') \sim \mu_t$ on both sides, we obtain:

$$\mathbb{E}_{\tilde{w}_t \sim \widetilde{p}_t} \left[\mathbb{E}_{\mathcal{S}} \left\|\nabla \widetilde{f}(\tilde{w}_t, \mathcal{S})\right\|_2^2\right] \leq 4L_\ell [Q(\widetilde{p}_t) - Q(p)] + 2\mathbb{E}_{(\tilde{w}_t, w') \sim \mu_t} \left[\mathbb{E}_{\mathcal{S}} \left\|\nabla \widetilde{f}(w', \mathcal{S})\right\|^2\right].$$

Therefore,

$$\begin{aligned}\mathbb{E}_{(\tilde{w}_t, w') \sim \mu_t} \left[\mathbb{E}_{w_t|\tilde{w}_t} \|w_t - w\|_2^2\right] \leq\ & \mathbb{E}_{(\tilde{w}_t, w') \sim \mu_t} \left[\|\tilde{w}_t - w\|_2^2\right] - 2\eta_t \left(f(\widetilde{p}_t) - f(p)\right) \\ & + \eta_t^2 \left(4L_\ell [Q(\widetilde{p}_t) - Q(p)] + 2\mathbb{E}_{(\tilde{w}_t, w') \sim \mu_t} \left[\mathbb{E}_{\mathcal{S}} \left\|\nabla \widetilde{f}(w', \mathcal{S})\right\|^2\right]\right),\end{aligned}$$

leading to the final result that

$$\begin{aligned}W_2^2(p_t,p) &\leq \mathbb{E}_{(\tilde{w}_t,w')\sim\mu_t}\left[\mathbb{E}_{w_t|\tilde{w}_t}\|w_t-w\|_2^2\right]\\ &\leq W_2^2(\widetilde{p}_t,p)-2\eta_t\left(f(\widetilde{p}_t)-f(p)\right)\\ &+\eta_t^2\left(4L_\ell[Q(\widetilde{p}_t)-Q(p)]+2\mathbb{E}_{(\tilde{w}_t,w')\sim\mu_t}\left[\mathbb{E}_{\mathcal{S}}\left\|\nabla\widetilde{f}(w',\mathcal{S})\right\|^2\right]\right).\end{aligned}$$

■

Proof [Proof of Lemma 19] Similar to the proof of Lemma 15, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\left\|\nabla\widetilde{f}(w',\mathcal{S})\right\|^2 \leq& 2\mathbb{E}_{\mathcal{S}}\|\nabla\widetilde{f}(w',\mathcal{S})-\nabla\widetilde{f}(w_*,\mathcal{S})\|_2^2+2\mathbb{E}_{\mathcal{S}}\|\nabla\widetilde{f}(w_*,\mathcal{S})\|_2^2\\ \leq& 4L_\ell\left[f(w)-f(w_*)-\nabla f(w_*)^\top(w-w_*)\right]+2\mathbb{E}_{\mathcal{S}}\|\nabla\widetilde{f}(w_*,\mathcal{S})\|_2^2.\end{aligned}$$

Taking expectation on both sides, we obtain that

$$\begin{aligned}\mathbb{E}_{(\tilde{w}_t,w')\sim\mu_t}\left[\mathbb{E}_{\mathcal{S}}\left\|\nabla\widetilde{f}(w',\mathcal{S})\right\|^2\right] \leq& 4L_\ell\mathbb{E}_{w\sim p}\left[f(w)-f(w_*)-\nabla f(w_*)^\top(w-w_*)\right]\\ &+2\mathbb{E}_{\mathcal{S}}\left\|\nabla\widetilde{f}(w_*,\mathcal{S})\right\|^2.\end{aligned}\tag{32}$$

We now upper bound $\mathbb{E}_{w\sim p}\left[f(w)-f(w_*)-\nabla f(w_*)^\top(w-w_*)\right]$. Let p_0 be the normal distribution $N(w_*,(\beta/m)I)$, and define

$$\Delta f(w)=f(w)-f(w_*)-\nabla f(w_*)^\top(w-w_*),\tag{33}$$

$$\Delta g(w)=g(w)-g(w_*)-\nabla g(w_*)^\top(w-w_*),\tag{34}$$

then p can be expressed as

$$p\propto\exp(-\beta^{-1}(\Delta f(w)+\Delta g(w))),$$

which is the solution of

$$p=\arg\min_p\mathbb{E}_{w\sim p}\left[\Delta f(w)+\beta\ln\frac{p(w)}{p_0(w)}\right].$$

Therefore

$$\begin{aligned}\mathbb{E}_{w\sim p}\Delta f(w) \leq&\mathbb{E}_{w\sim p}\left[\Delta f(w)+\beta\ln\frac{p(w)}{p_0(w)}\right]\\ =&-\beta\ln\mathbb{E}_{w\sim p_0}\exp(-\beta^{-1}\Delta f(w))\\ \leq&-\beta\ln\mathbb{E}_{w\sim p_0}\exp(-0.5\beta^{-1}(w-w_*)^\top H(w-w_*))\\ =&0.5\beta\ln|I+H/m|\leq 0.5\beta\mathrm{trace}(H/m).\end{aligned}\tag{35}$$

Since $\mathbb{E}_{\mathcal{S}}\nabla\widetilde{f}(w_*,\mathcal{S})=\nabla f(w_*)$, we can decompose $\mathbb{E}_{\mathcal{S}}\left\|\nabla\widetilde{f}(w_*,\mathcal{S})\right\|^2$ as follows:

$$\mathbb{E}_{\mathcal{S}}\left\|\nabla\widetilde{f}(w_*,\mathcal{S})\right\|^2=\|\nabla f(w_*)\|^2+\mathbb{E}_{\mathcal{S}}\left\|\nabla f(w_*)-\nabla\widetilde{f}(w_*,\mathcal{S})\right\|^2.$$

Since w_* is the minimum of $f+g$, $\nabla f(w_*) = -\nabla g(w_*) = -mw_*$. Hence

$$\mathbb{E}_{\mathcal{S}} \left\| \nabla \widetilde{f}(w_*, \mathcal{S}) \right\|^2 = m^2 \left\| w_* \right\|^2 + \mathbb{E}_{\mathcal{S}} \left\| \nabla f(w_*) - \nabla \widetilde{f}(w_*, \mathcal{S}) \right\|^2 . \tag{36}$$

Plugging inequalities (35) and (36) into inequality (32) proves the desired result that

$$\begin{aligned}&\mathbb{E}_{(\tilde{w}_t, w') \sim \mu_t} \left[\mathbb{E}_{\mathcal{S}} \left\| \nabla \widetilde{f}(w', \mathcal{S}) \right\|^2 \right] \\ \leq &\frac{2\beta L_\ell}{m} \mathrm{trace}(H) + 2m^2 \left\| w_* \right\|^2 + 2\mathbb{E}_{\mathcal{S}} \left\| \nabla f(w_*) - \nabla \widetilde{f}(w_*, \mathcal{S}) \right\|^2 .\end{aligned}$$

■

Acknowledgments

This work is supported in part by the National Science Foundation Grants NSF-SCALE MoDL(2134209) and NSF-CCF-2112665 (TILOS), the U.S. Department of Energy Office of Science, and the Facebook Research Award.

References

Alekh Agarwal, Peter L. Bartlett, Pradeep D. Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58:3235–3249, 2012.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2nd edition, 2008.

D. Bakry and M. Emery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pages 177–206. 1985.

E. Bernton. Langevin Monte Carlo and JKO splitting. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 1777–1798, 2018.

N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. arXiv:1805.00452, 2018.

G. E. P. Box and G. C. Tiao. *Bayesian inference in statistical analysis*. John Wiley & Sons, 1992.

N. Chatterji, N. Flammarion, Y.-A. Ma, P. Bartlett, and M. Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 764–773, 2018.

Niladri S. Chatterji, Jelena Diakonikolas, Michael I. Jordan, and Peter L. Bartlett. Langevin monte carlo without smoothness. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, volume 108, 2020.

X. Cheng and P. L. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, pages 186–211, 2018.

X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. arXiv:1805.01648, 2018a.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 300–323, 2018b.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. Royal Stat. Soc. B*, 79(3):651–676, 2017.

A. S. Dalalyan and A. G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. arXiv:1710.00095, 2017.

A. S. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. arXiv:1807.09382, 2018.

John C. Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 06 2017.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! arXiv:1801.02309, 2018.

J. B. Gelman, A.and Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, New York, 2004.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15: 2489–2512, 2014.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, January 1998.

Simon Lacoste-Julien, Mark Schmidt, and Francis R. Bach. A simpler approach to obtaining an o(1/t) convergence rate for the projected stochastic subgradient method. *CoRR*, abs/1212.2002, 2012.

G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133: 365–397, 2012.

M. Ledoux. The geometry of Markov diffusion generators. *Ann Fac Sci Toulouse Math*, 9(6): 305–366, 2000.

Y.-T. Lee, Z. Song, and S. S. Vempala. Algorithmic theory of ODEs and sampling from well-conditioned logconcave densities. arXiv:1812.06243, 2018.

Q. Li and N. Lin. The Bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 2010.

Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *Proc. Natl. Acad. Sci. U.S.A.*, 116:20881–20885, 2019.

Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for MCMC? *Bernoulli*, 27(3):1942–1992, 2021.

O. Mangoubi and A. Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. arXiv:1708.07114, 2017.

O. Mangoubi and N. K. Vishnoi. Dimensionally tight running time bounds for second-order Hamiltonian Monte Carlo. arXiv:1802.08898, 2018.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

W. Mou, Y.-A. Ma, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. High-order Langevin diffusion yields an accelerated MCMC algorithm. *Journal of Machine Learning Research*, 22:1–41, 2021.

Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. An efficient sampling algorithm for non-smooth composite potentials. arXiv:1910.00551, 2019.

R. M. Neal. *Bayesian Learning for Neural Networks*. Number 118 in Lecture Notes in Statistics. Springer-Verlag, New York, 1996.

F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, 173(2):361–400, 2000.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. 30 (4):838–855, 1992.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International conference on machine learning*, pages 1571–1578, 2012.

G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.*, 4:337–357, 2002.

P. J. Rossky, J. D. Doll, and H. L. Friedman. Brownian dynamics as smart Monte Carlo simulation. *J. Chem. Phys.*, 69(10):4628, 1978.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79. PMLR, 2013.

R. Shen and Y. T. Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2098–2109, 2019.

P. Sollich. Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning*, 46:21–52, 2002.

S. S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Log-Sobolev suffices. arXiv:1903.08568, 2019.

C. Villani. *Optimal Transport, Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009.

A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 2093–3027, 2018.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

Difan Zou and Quanquan Gu. Stochastic gradient hamiltonian monte carlo methods with recursive variance reduction. In *Advances in Neural Information Processing Systems (NeurIPS) 32*. 2019.